UNIVERSIDAD DE SALAMANCA

DEPARTAMENTO DE ESTADÍSTICA

NUEVAS APROXIMACIONES METODOLÓGICAS AL ESTUDIO DE LA COLABORACIÓN EN LA CIENCIA A TRAVÉS DE LAS PUBLICACIONES CIENTÍFICAS

TESIS DOCTORAL

ADRIÁN ARIAS DÍAZ-FAES

2015

NUEVAS APROXIMACIONES METODOLÓGICAS AL ESTUDIO DE LA COLABORACIÓN EN LA CIENCIA A TRAVÉS DE LAS PUBLICACIONES CIENTÍFICAS

Memoria que para optar al Grado de Doctor, por el Departamento de Estadística de la Universidad de Salamanca, presenta:

Adrián Arias Díaz-Faes

Salamanca

2015

Esta tesis doctoral corresponde a un compendio de cuatro artículos previamente publicados o aceptados para su publicación en revistas científicas indexadas en el *Journal Citation Reports*:

Díaz-Faes, A.A.[1], González-Albo, B.[2], Galindo, M.P.[3], & Bordons, M.[1] (2013). HJ-Biplot como herramienta de inspección de matrices de datos bibliométricos. *Revista Española de Documentación Científica*, *36*(1): e001. doi: http://dx.doi.org/10.3989/redc.2013.1.988

Díaz-Faes, A.A.[1], & Bordons. M.[1] (2014). Acknowledgments in scientific publications: presence in Spanish science and text patterns across disciplines. *Journal of the Association for Information Science and Technology*, *65*(9), 1834-1849. doi: 10.1002/asi.23081

Bordons, M.[1], Aparicio, J.[2], González-Albo, B.[2], & Díaz-Faes, A.A.[1] (2015). The relationship between the research performance of scientists and their position in co-authorship networks in three fields. *Journal of Informetrics*, *9*(1), 135-144. doi: 10.1016/j.joi.2014.12.001

Díaz-Faes, A.A.[1], Costas, R.[4], Galindo, M.P.[3], & Bordons, M.[1] (en prensa[*]). Unravelling the performance of individual scholars: use of Canonical Biplot analysis to explore the performance of scientists by academic rank and scientific field. *Journal of Informetrics*.

Afiliación de los autores:

[1] Departamento de Ciencia, Tecnología y Sociedad, IFS, Consejo Superior de Investigaciones Científicas (CSIC), Albasanz 26-28, 28037 Madrid (España).

[2] Centro de Ciencias Humanas y Sociales (CCHS), Consejo Superior de Investigaciones Científicas (CSIC), Albasanz 26-28, 28037 Madrid (Spain).

[3] Departamento de Estadística, Universidad de Salamanca, Alfonso X El Sabio s/n, 37007 Salamanca (España).

[4] Centre for Science and Technology Studies (CWTS), Universidad de Leiden, PO Box 905 2300 AX Leiden (Países Bajos).

[*] *Se adjunta a continuación carta de aceptación de la revista.*

CARTA DE ACEPTACIÓN

Ref: JOI_2015_20

Status: Accept, Date: 17/Apr/2015 14:22

Title: Unravelling the performance of individual scholars: use of Canonical Biplot analysis to explore the performance of scientists by academic rank and scientific field

Journal of Informetrics

Dear Mr. A. Díaz-Faes,

I am pleased to inform you that your paper Unravelling the performance of individual scholars: use of Canonical Biplot analysis to explore the performance of scientists by academic rank and scientific field meets the standards of Journal of Informetrics and has been accepted for publication.

My comments and those of the reviewers are appended at the end of this message.

Thank you for submitting your work to Journal of Informetrics.

When your paper is published on ScienceDirect, you want to make sure it gets the attention it deserves. To help get your message across, Elsevier has developed a free new service called AudioSlides: brief, webcast-style presentations that are shown (publicly available) next to your published article. This format gives you the opportunity to explain your research in your own words and attract interest. You will receive an invitation email to create an AudioSlides presentation shortly. For more information and examples, please visit http://www.elsevier.com/audioslides.

Kind regards,

Vincent Larivière

Associate Editor

Journal of Informetrics

Editor and reviewer comments:

The authors have appropriately addressed the minor revisions requested by the two referees.

DEPARTAMENTO DE ESTADÍSTICA

MARÍA BORDONS GANGAS

Profesora de Investigación del Departamento de Ciencia, Tecnología y Sociedad, IFS, Consejo Superior de Investigaciones Científicas

Y

Mª PURIFICACIÓN GALINDO VILLARDÓN

Profesora Titular del Departamento de Estadística de la Universidad de Salamanca

CERTIFICAN: Que **D. Adrián Arias Díaz-Faes** ha realizado, bajo su dirección, el trabajo que, para optar el Grado de Doctor, presenta, mediante compendio de publicaciones, con el título: *Nuevas aproximaciones metodológicas al estudio de la colaboración en la ciencia a través de las publicaciones científicas*, y para que conste, firman el presente certificado en Salamanca, en junio de 2015.

Fdo.: María Bordons Gangas

Fdo.: Mª Purificación Galindo Villardón

VNiVERSiDAD
Đ SALAMANCA

Departamento de Estadística
Facultad de Medicina
C/ Alfonso X el Sabio s/n 37007 SALAMANCA
Tfno.: 923 294 400 Ext. 1921 Fax: 923 294 619
Email: sestadistica@usal.es

Dª Mª Purificación Galindo Villardón, Coordinadora del Programa de Doctorado de la Universidad de Salamanca *Estadística Multivariante Aplicada* y Presidenta de su Comisión de Docencia, **INFORMA:**

Que en la reunión de la Comisión de Docencia de 24 de junio de 2015 se acordó por unanimidad emitir informe favorable a la solicitud de depósito de la tesis presentada por **D. Adrián Arias Díaz-Faes** en la modalidad de Compendio de Artículos.

Y para que surta los efectos oportunos, expido y firmo el presente en Salamanca, a 24 de junio de 2015.

LA PRESIDENTA DE LA COMISIÓN DE DOCENCIA

Fdo.: Mª Purificación Galindo Villardón

We shall not cease from exploration

And the end of all our exploring

Will be to arrive where we started

And know the place for the first time.

— T.S. Eliot, Four Quartets

Agradecimientos

En un sentido formal, este trabajo ha sido posible gracias a la concesión de una beca/contrato del programa de la Junta para la Ampliación de Estudios (JAE predoc 2011) del CSIC para el desarrollo de tesis doctorales. Además, hay una serie de personas que, de una u otra manera, han contribuido de forma sustancial a la consecución de esta tesis doctoral.

En primer lugar, quiero expresar mi admiración a mi directora María Bordons con quien ha sido un privilegio poder trabajar durante estos años y quien me ha enseñado el oficio de investigador. Sus conocimientos, dedicación, rigurosidad y capacidad de análisis, de la que aún me sorprendo, han hecho que muchas de las ideas que han sobrevolado por mi cabeza se hayan podido plasmar en trabajos científicos. A mi co-directora Puri Galindo a quién conocí como alumno y me abrió un camino inesperado y apasionante. Has sido una mentora incansable a lo largo de estos años. Gracias por tu confianza, cercanía y entrega.

En segundo lugar, quiero reconocer el apoyo y amistad de los miembros que son o han sido parte del grupo ACUTE. En especial, gracias a Javier Aparicio, Borja González-Albo, Luz Moreno y Nacho Santabárbara, sin vuestro trabajo previo y asistencia los trabajos que aquí figuran no habrían sido posibles. A Isabel Gómez y Nana Morillo por sus consejos, ayuda y sugerencias.

También me gustaría expresar mi gratitud a Birger Larsen de la Universidad de Aalborg por su apoyo durante la etapa inicial de esta tesis doctoral. Asimismo, quiero agradecer a Rodrigo Costas y Thed van Leeuwen del *Centre for Science and Technology Studies (CWTS)* a quienes admiro por sus conocimientos y entusiasmo, y quienes, junto a otras colegas (Clara Calero, Alfredo Yegros, Ludo Waltman), hicieron de mi estancia en la Universidad de Leiden algo inolvidable, tanto en lo profesional como en lo personal.

Finalmente, a mi madre, por no haber perdido nunca la confianza en mí y por ser siempre una fuente de apoyo. Y como no, a Nuria por haber sido mi compañera de viaje durante todos estos años.

ÍNDICE

Índice

Summary 1

Justificación y estructura de la tesis 13

PARTE 1. ANTECEDENTES, METODOLOGÍA Y CONCLUSIONES 15

1. Introducción 17

1.1. La medición de la ciencia 17

1.2. El fenómeno de la colaboración 22

1.3. La multidimensionalidad de la ciencia 28

1.4. ¿Por qué una aproximación multivariante? 31

1.5. Análisis multivariante en bibliometría 34

2. Conceptos metodológicos 39

2.1. La fuente de datos: *Web of Science* 39

2.2. Tratamiento y normalización de los datos 43

2.3. Niveles de análisis y consideraciones generales 46

2.4. Clasificación de los indicadores bibliométricos 50

2.4.1. Indicadores según la aplicación de estándares de referencia 50

2.4.2. Indicadores según las dimensiones de la actividad científica 52

3. Asunciones de partida y preguntas de investigación 69

4. Conclusiones generales e investigación futura 71

Referencias 76

PARTE 2. ARTÍCULOS PUBLICADOS 87

5. Resumen de las publicaciones 89

6. HJ-Biplot como herramienta de inspección de matrices de datos bibliométricos 95

6.1. Introducción 95

6.2. Material y métodos 97

6.2.1. Objeto de estudio 97

6.2.2. Métodos Biplot 99

6.3. Resultados 104

6.3.1. Análisis del impacto y la colaboración: plano 1-2 105

6.3.2. Análisis del impacto y la colaboración: plano 1-3.....108
6.3.3. Clusters según tipo de colaboración.....110
6.4. Discusión y conclusiones 112
Referencias 115
Anexo. Relación de centros propios y mixtos del CSIC por áreas científico-técnicas (2006-2009) 118
7. Unravelling the performance of individual scholars: use of Canonical Biplot analysis to explore the performance of scientists by academic rank and scientific field 121
7.1. Introduction 121
7.2. Methods.....124
7.2.1. Data and bibliometric indicators 124
7.2.2. Canonical Biplot 128
7.2.3. Effect sizes.....130
7.3. Results.....130
7.3.1. Biplot analysis 131
7.3.2. Effect size results 137
7.4. Discussion 138
References 141
8. The relationship between the research performance of scientists and their position in co-authorship networks in three fields 145
8.1. Introduction 145
8.2. Research questions.....148
8.3. Methods.....149
8.3.1. Social network measures 149
8.3.2. Measures of research performance 152
8.4. Results.....153
8.4.1. Network structure.....153
8.4.2. Relationship between performance indicators and the position of authors in networks.....155
8.5. Discussion 159
References 162

9. Acknowledgments in scientific publications: presence in Spanish science and text patterns across disciplines.....165
9.1. Introduction.....165
9.2. Objectives.....169
9.3. Data and methods.....169
9.3.1. Data sources.....170
9.3.2. Analysis of FA presence by subject area.....170
9.3.3. Analysis of textual patterns in four subject categories.....171
9.4. Results.....174
9.4.1. Analysis of FA presence by subject area.....175
9.4.2. Influence of journal prestige.....176
9.4.3. Influence of the number of authors.....179
9.4.4. Influence of the research level.....181
9.4.5. Analysis of textual patterns by subject category.....181
9.5. Discussion.....184
9.5.1. Funding acknowledgment by subject area.....185
9.5.2. Funding acknowledgment by journal prestige.....186
9.5.3. Funding acknowledgment by number of authors.....187
9.5.4. Acknowledgment patterns by subject category.....188
9.5.5. Authors, subauthors and contributors.....189
9.5.6. Limitations of the study.....190
9.6. Conclusions and future research.....190
References.....192

ANEXOS.....197
I. Áreas y disciplinas *Web of Science*.....199
II. Fundamentos teóricos del HJ-Biplot.....202
III. Fundamentos teóricos del Biplot Canónico.....205
IV. Índices de calidad de las publicaciones aportadas.....208

Summary

Introduction

Science has been one of the most important social and cultural phenomena over the last five centuries, and in particular, since World War II. The social import of science and its intrinsic complex mechanisms explain the growing interest in understanding the different underlying processes and dynamics at play. Although traditionally scientific progress has been associated with individual ventures, the truth is that science is not an individual endeavour but a collaborative enterprise (Merton, 1968). Nowadays, collaboration plays an increasingly essential role in scientific activity and is deemed to be one of the hallmarks of modern science (Bordons & Gómez, 2000; Wren et al., 2007; Gazni, Sugimoto & Didegah, 2012).

Collaboration enables scientists to share knowledge, expertise and techniques; cope with interdisciplinary research topics; and get involved in sophisticated research projects (Katz & Martin, 1997). Because of these important functions and considering the limited availability of resources and the increasing costs of research, it is no wonder that collaborative research relies on funding agencies for financial support. At EU level, collaboration between teams from different Member States is a requisite in research projects developed under the framework program, as well as in networks of excellence to promote cohesion within the EU. In addition, the mobility of researchers is supported to foster international links. In Spain, the objectives of national and regional programs include the promotion of co-operation at different levels: between individuals, by encouraging the creation of research groups; across disciplines, by promoting interdisciplinary research; across institutional sectors, by fostering contacts between universities and the private sector; and between centres, by creating networks, particularly in the biomedicine field.

Bibliometrics' quantitative approach to the study of science and scientific collaboration has been historically linked to the idea that it is possible to convey human knowledge through publications and their components. As stated by Martin and Irvine (1983), science may be conceived as an input-output process. Its outputs may be classified according to their nature under different dimensions (Martin, 1996), namely, the scientific dimension (contributions to scientific expertise), the educational dimension (contributions to the development of training and skills), the technological dimension (contributions to the development and improvement of technologies) and, last but not least, the cultural dimension (contributions to society at large). The basic assumption of bibliometrics is that the scientific dimension, i.e. the creation of new knowledge, plays a key role in science. As new knowledge only acquires value when published, scientific publications are the cornerstone of science and its results (Kostoff, 1995).

The classical bibliometric approach to the study of scientific collaboration is based on the analysis of co-authorship, which provides relevant information on the structure and dynamics of collaboration between researchers. However, as noted in the literature, if we reduce collaboration to co-authorship some collaborative activity may be overlooked (Laudel, 2002; Melin & Persson, 1996). Consequently, it is interesting to study other sources of information on collaboration such as the acknowledgments section in scientific articles, the usefulness of which as a measure of "sub-authorship collaboration" has been suggested elsewhere in the literature (Patel, 1973).

To measure the true extent of collaboration it is necessary to bear in mind its interactions and repercussions on different aspects of research activity. Accordingly, several studies have shown the positive influence of collaboration on impact, productivity or interdisciplinarity. Thus, the stronger impact of multi-authored publications as compared to single-authored ones has been widely described (Persson, Glänzel, & Danell, 2004), as well as the upward trend of impact along with the number of authors and institutions, and wider scope of collaboration (Bordons, Aparicio, & Costas, 2013; Abramo, D'Angelo, & Solazzi, 2011). Special benefits have been found to derive from international collaboration (Franceschet & Costantini, 2010).

Other studies have described a strong relationship between collaboration and scientific productivity (Pao, 1992). However, the association between these variables depends on the disciplines involved (Abramo, D'Angelo, & Di Costa, 2009) due not only to the different nature of their research, but also to the fact that collaboration implies additional costs, as those arising from the need to co-ordinate research (Cummings & Kiesler, 2005). Such costs should not exceed the benefits derived from the research to make collaboration worthwhile for researchers.

As regards interdisciplinarity, research is increasingly becoming an activity requiring interaction between a number of distinct disciplines (Porter & Rafols, 2009). Collaboration may contribute to interdisciplinarity since it involves participants with different specialisation profiles and from different organisational contexts, thus prompting a higher degree of heterogeneity in research-team composition (Bordons et al., 2013). Partner diversity contributes to boost the benefits of collaboration since it brings into play a wider variety of viewpoints leading to higher creativity (Reagans & Zuckerman, 2001).

Collaboration in science has been studied in the literature at different levels of analysis, ranging from studies by countries (macro level) (Glänzel, 2001), by disciplines or institutions (meso level) (Francescht & Costantini, 2010), and by teams or individual scientists (micro level) (see for example He, 2009). Micro-level studies are particularly interesting and have received special attention in the last few years mainly due to the key role played by bibliometric indicators in research assessment processes. At this level, special caution is required as regards the collection of data and the calculation of

indicators bearing in mind the difficulties that must be sorted out for the correct identification of the entire production of researchers and the lower validity of statistical analyses applied to small units. Nonetheless, once these difficulties have been overcome, studies at the micro level may prove especially revealing since bibliometric data may be complemented with organisational and personal factors including, but not limited to, age, gender and professional category, allowing for the conduct of more comprehensive studies (Costas, van Leeuwen, & Bordons, 2010).

On the basis of the foregoing, it becomes clear that collaboration is a complex phenomenon which may influence different aspects of scientific activity. In this context, multivariate analysis emerges as a significant tool with an important potential for the development of bibliometrics. On the one hand, it enables us to outline and describe the complex underlying structure of data and the interactions between different bibliometric indicators. On the other hand, bibliometric data have a series of features, such as asymmetric distributions (Seglen, 1992), highly-correlated variables, or large differences in sample sizes, that hinder the implementation of conventional statistical methods (macro-level samples may comprise millions of documents, while micro-level ones may be very small). As a result, during the last decade, a growing use of multivariate analysis techniques has been highly noted in bibliometrics, chiefly applying Multidimensional Scaling, Factor Analysis based on Principal Components Analysis, and Cluster Analysis.

Besides, this dissertation makes the basic assumption that scientific activity has a multidimensional nature (Martin & Irvine, 1983). This means that science must be understood as a social phenomenon in which all the elements involved interact and where it is not possible to isolate the impact of some factors from that of other factors; nor to provide an unambiguous description thereof (Moravsick, 1984). Therefore, a diverse set of variables and methodologies need to be considered to provide a clear picture of research activity. In this context, new methodological approaches are introduced from the perspective of multivariate statistics, especially through Biplot methods and social network analysis. In addition, due examination is made of the role of collaboration in science at large and its relationship with other variables concerning different aspects of research and personal factors, different levels of aggregation (meso and micro) are taken into account, and the differences across scientific areas are emphasised.

Objectives and research questions

This dissertation, which comprises four articles in refereed journals (three of them already published and one pending approved publication), introduces new methodological approaches to the study of collaboration in science and aims to further

our insight on this complex phenomenon. The following questions are specifically addressed:

- ✓ *What new avenues do Biplot methods open up for the study of collaboration in science at the meso (centres) and micro level (individuals)?*

  An assessment of the usefulness of Biplot methods in bibliometric studies is conducted. Assuming the multidimensional nature of research, the existence of interactions among different indicators of research performance and the benefits of obtaining an integrated graphical representation of results, the HJ-Biplot is deemed to be an interesting tool for bibliometricians. HJ-Biplot is applied to outline the scientific activity of CSIC's research centres allowing for the simultaneous interpretation of the interaction between the indicators retained for the purposes of this study and the position of the centres according to their performance (Publication 1). At the micro level, Canonical Biplot is used to explore the research performance of Spanish researchers at the CSIC grouped by field and academic rank through a diverse set of bibliometric indicators. This technique enables us to build a Biplot where the groups of individuals are sorted out by the maximum discriminating power between the different selected indicators (Publication 2).

- ✓ *What is the contribution of social network analysis to the study of collaboration in scientific publications? What is the relationship between the research performance of scientists and their position in co-authorship networks?*

  Social network analysis may be used for the study of collaboration among different units and may provide diagrams which represent not only the structure of a field, but also network measures for the characterisation of the whole system and positional measures to describe each of the units. The structure of co-authorship networks in three different fields in Spain is examined by means of centrality and cohesion measures (Publication 3). A general description of the networks is laid out and the behaviour of authors on the basis of their relationships with other authors is analysed. In addition, the relationship between the research performance of scientists and their position in co-authorship networks is explored as well.

- ✓ *Can we trace sub-authorship collaboration through the study of the acknowledgments section of papers? Are there differences by discipline in the presence of acknowledgments and in the nature of collaboration?*

  Assuming that the acknowledgements section of a paper is an interesting part of scholarly communication because it facilitates the collection of data on special contributions to research that are not rewarded with authorship, an appraisal of the potentialities of said sections as a source of sub-authorship collaboration is

presented (Publication 4). The fact that the Web of Science has been including funding acknowledgment data since August 2008 opens up new possibilities for data mining and the analysis of the information contained in the acknowledgments section of papers. Publication 4 analyses the presence of acknowledgments in English language papers published by Spanish researchers by subject area and proposes a novel approach to reveal sub-authorship information through text mining and Correspondence Analysis.

Publications

1) HJ-Biplot as a tool for inspection of bibliometric data matrices. *Revista Española de Documentación Científica*, 36(1). e001. doi: http://dx.doi.org/10.3989/redc.2013.1.988

Bibliometric studies incorporate increasingly sophisticated indicators, advanced statistical techniques and visualisation tools. Even though visualisation tools have recently experienced a remarkable surge well above that observed for multivariate techniques, it is the latter that bear an enormous potential for bibliometrics.

The aim of this paper is to demonstrate the usefulness of HJ-Biplot in bibliometric studies. Biplot methods are graphical representations of multivariate data. Using HJ-Biplot it is possible to interpret simultaneously the position of the centres, represented by dots; indicators, represented by vectors; and the relationships between them. It provides a simple and intuitive display, similar to a scatterplot, but capturing the multivariate co-variance structures between bibliometric indicators. Their interpretation does not require specialised statistical knowledge, but merely to know how to interpret the length of a vector, the angle between two vectors and the distance between two points. With this aim, an analysis of the scientific output of CSIC's own centres as well as of joint centres during the period 2006-2009 has been conducted using a series of indicators based on impact (percentage of Q1 articles, normalised position, and relative world citations) and collaboration (percentage of non-collaborative papers, percentage of national collaboration, and percentage of international collaboration).

The outcomes show a positive correlation between impact indicators and international collaboration. Research performance is partly dependent on each area, since research centres and institutes within any given area tend to be placed close to each other in the plot. Nonetheless, there is also some intra-area heterogeneity. In fact, humanities and social sciences, and food science and technology show higher homogeneity, while this is found to be lower among research centres for physics and agricultural sciences.

In addition, it is possible to identify research centres with an outstanding or unusual behaviour within each single area.

Among the advantages of the HJ-Biplot analysis we find that it is applicable to any data matrix and not only frequencies, unlike other multivariate techniques such as Correspondence Analysis, which also allows to obtain a simultaneous plot for rows and columns in a reduced dimensional space

2) Unravelling the performance of individual scholars: use of Canonical Biplot analysis to explore the performance of scientists by academic rank and scientific field. *Journal of Informetrics* (in press).

There is a growing interest in bibliometric studies at the individual level, where it is important to consider not only the dimensions of scientific activity (collaboration, impact, level of research, interdisciplinarity), but also personal and academic factors which may influence the scientific performance of researchers. In this context, the use of multivariate analysis techniques can be particularly relevant.

In this paper, the Canonical Biplot technique is introduced to explore differences in the scientific performance of Spanish CSIC researchers between 2007-2011, organised by field (Chemistry and Materials Science) and grouped by academic rank (research fellows and three types of full-time permanent scientists). This method enables us to build a Biplot where the groups of individuals are sorted out by the maximum discriminating power between the different indicators considered. Besides, as confidence intervals are displayed in the plot, statistical differences between groups are liable to be studied simultaneously. Since test hypotheses are sensitive to different sample size effects, sizes for some pairwise comparisons are computed by means of Hedge's *g*.

Two gradients for CSIC scientists are observed. First, researchers of different academic ranks are clearly separated according to their age, level of production of papers, distinct number of collaborators, the number of highly-cited papers and their position in the byline. The second gradient relates to intrinsic field features since it separates Chemistry from Materials Science. Size effects support the outcomes found.

In contrast to other multivariate techniques, Canonical Biplot offers some interesting features. For instance, if we would have applied a MANOVA, we should have examined many tables and we would not have obtained a joint representation in a low dimensional space for a visual inspection of the underlying structure of the data matrix. If we would have used a Discriminant Analysis, we would have obtained a low dimensional plot describing the group's structure, but we would not have had direct information about the bibliometric indicators responsible for the separation between

groupings and their correlations. Moreover, effect sizes can be a relevant measure at the individual-level since data collection at this level is a tough job and it is not always possible to achieve a good sample size for the different categories or groupings.

It is concluded that the Canonical Biplot analysis is a strong exploratory tool with high potential in order to make headway in the unravelling of the intricate structure of relationships between research performance indicators and the individual characteristics of researchers.

3) The relationship between the research performance of scientists and their position in co-authorship networks in three fields. *Journal of Informetrics*, *9*(1), 135-144. doi: 10.1016/j.joi.2014.12.001

Research networks play a crucial role in the production of new knowledge since collaboration contributes to determine the cognitive and social structure of scientific fields and has a positive influence on research. In recent times, social network analysis has emerged as an appealing approach to the study of co-authorship in science. This approach enables to deepen in the dynamics of scientific production by discipline, link certain collaboration practices to better performance and identify authors who hold strategic positions within networks.

This paper analyses the structure of co-authorship networks in three different fields (nanoscience, pharmacology and statistics) in Spain over a three-year period (2006-2008) and explores the relationship between the research performance of scientists and their position in co-authorship networks. A number of indicators to measure scientific activity (number of articles, citations, g-index) and several social network measures related to centrality (degree centrality, closeness centrality, betweenness centrality, centralisation, eigenvector centrality) and cohesion (strength of ties, constraint, clustering coefficient) are used.

At the macro level, a denser co-authorship network is found in the two experimental fields (pharmacology and nanoscience) than in statistics, where a less connected and more fragmented network is observed (high constraint and lower propensity of authors to form cliques). Likewise, the main component includes around two thirds of the authors in the denser networks (pharmacology and nanoscience), as against only 28% in statistics.

Using the g-index as a proxy for individual research performance, a Poisson regression model is used to explore how performance is related to different co-authorship network measures and to disclose inter-field differences. The study at the micro-level confirms that there is a relationship between the position of Spanish scientists in co-

authorship networks and his/her research performance as measured by the g-index. This association varies by field and seems to be stronger in pharmacology and nanoscience than in statistics. The number of co-authors (degree centrality) and the strength of links show a positive relationship with the g-index in the three fields. Local cohesiveness presents a negative relationship with g-index in the two experimental fields, where open networks and the diversity of co-authors seem to be beneficial. No clear advantages from intermediary positions (high betweenness) or from being linked to well-connected authors (high eigenvector) can be inferred from this analysis. In terms of g-index, the benefits derived by authors from their position in co-authorship networks are larger in the two experimental fields than in the theoretical one.

4) Acknowledgments in scientific publications: presence in Spanish science and text patterns across disciplines. ***Journal of the Association for Information Science and Technology*****,** ***65*****(9), 1834-1849. doi: 10.1002/asi.23081**

Acknowledgments have become an important feature in scholarly communication, since they are used to recognize some special contributions to research that do not qualify for authorship status (sub-authorship), but may well have a significant bearing on the final results of research. In fact, the scientific literature has pointed out their cognitive, social and instrumental meaning. Until recently, it was very difficult to carry out studies on acknowledgments, because this information was not available in bibliographic databases. However, the Web of Science has been including funding acknowledgment data since August 2008, which opens up new possibilities for data mining and the analysis of the information contained in the acknowledgment section of papers.

This research aims to increase our knowledge about the presence of the acknowledgments in scientific publications and explores its usefulness as a source of information on scientific collaboration. First, the presence of acknowledgments in 38,257 English language papers published by Spanish researchers in 2010 is studied by subject area on the basis of the funding acknowledgment information available in the Web of Science database. Second, a novel approach for discovering text patterns by discipline in the acknowledgment section of papers is introduced by means of text mining and Correspondence Analysis.

Funding acknowledgments are present in two thirds of Spanish articles, with significant differences by subject area (lower frequency in social sciences and humanities), and by basic/applied nature of research (in clinical medicine) and showing a higher presence in high-impact-factor journals. In addition, a higher number of authors per paper are found for those papers with funding acknowledgments, which is often related to a more complex and basic research. The existence of specific acknowledgment patterns

in English-language papers of Spanish researchers in four selected disciplines (cardiac and cardiovascular systems, economics, evolutionary biology, and statistics and probability) is revealed. “Peer interactive communication” predominates in the more theoretical or social-oriented fields (statistics and probability, economics), whereas the recognition of technical assistance is more common in experimental research (evolutionary biology), and the mention of potential conflicts of interest emerges forcefully in the clinical field (cardiac and cardiovascular systems). It becomes clear that the content of the acknowledgment section varies largely by discipline and contains very heterogeneous data that go beyond financial support and include sub-authorship information.

We conclude that the study of the acknowledgment section is an interesting option for an in-depth analysis of collaborative research practices, assuming that a sizeable part of them remains beyond the scope of the classical bibliometric indicators used to measure research collaboration (co-authorship). Besides, certain developments in the way in which acknowledgment information is included in the Web of Science may enhance future research on the topic. First and foremost, the collection of funding acknowledgment data for all journals, regardless of their language, would be desirable. Second, the inclusion of the acknowledgment section in all Web of Science records, and not only when funding is acknowledged, would allow more global, comprehensive, and accurate studies. Finally, the systematic inclusion of structured data about acknowledgments in journal articles and bibliographic databases would have a positive impact on the study of collaboration practices in science.

Conclusions

This research provides an assessment of the usefulness of different methodologies based on multivariate statistics and social network analysis for the study of collaboration in science. The information arising from the application of these approaches may be of interest not only for policymakers and research managers since the results obtained can be used advantageously in the decision-making process; but also for scholars from different disciplines interested in the study of the scientific process, and for scientists themselves.

With regard to Biplot methods, they emerge as powerful multivariate tools to be incorporated into bibliometric studies. Specifically, HJ-Biplot and Canonical Biplot have proved to adjust well to the special features of bibliometric data at different levels of analysis; moreover, they enable us to delve into the underlying data structure and to outline the relationships between the units of study and a diverse set of indicators. At the meso level, differences in the research performance of centres concerning production, collaboration and impact are shown by area and centre proving that an

outstanding or unique behaviour within each area can be identified. At the micro level, the research performance of scientists, including collaborative practices, is assessed along with its relationship with personal factors, such as academic rank and age. From the standpoint of scientific policy, international collaborative networks are found to be a suitable framework to boost research performance and lead science to higher levels of excellence, although future recommendations should bear in mind the specific features of each area.

On the other hand, the social network analysis approach, which includes the use of centrality and cohesion measures, offers valuable information not only for the description of the global structure of scientific fields, but also for characterising the position and role of every author in the network and their relationship with research performance. Interestingly, having a high number of collaborators and/or high strength of links with co-authors is associated with higher research performance, suggesting that these factors should be fostered, while no clear benefits of intermediary positions in the network are observed.

This research also reveals that the acknowledgements section can be an appealing source of information on collaboration issues. In this regard, the approach proposed has proved successful to extract sub-authorship information from the acknowledgments section of papers filed in the Web of Science by means of text mining and Correspondence Analysis, while interfield differences in the prevailing type of collaboration are also identified. However, a more comprehensive and better structured collection of information from the acknowledgements section of papers in the Web of Science would be required to facilitate automatic data processing.

Since literature has suggested that when we reduce collaboration to co-authorship we are running the risk of neglecting some collaborative activity (Melin & Persson, 1966; Laudel, 2002) incorporating information from the acknowledgements sections can be particularly forthcoming, while using advanced techniques, such as multivariate and social network analyses, may prove critical to unravel the complex phenomenon of collaboration in science. In this context, the importance of having an in-depth knowledge of multivariate analysis methods should be pointed out with the aim of identifying the most appropriate technique to be used in each case. This need was already identified almost three decades ago by Tijssen and de Leeuw (1988), but it has become far more significant now due to the increasing role played by multivariate analysis in bibliometrics studies and, in particular, in the so-called "maps of science" (which sometimes include multivariate methods) which have been proposed as a useful tool to support research management and scientific policy development (Noyons, 2001).

References

Abramo, G., D'Ángelo, C.A., & Di Costa, F. (2009). Research collaboration and productivity is there correlation? *Higher Education, 57*, 155-171.

Abramo, G., D'Angelo, C.A., & Solazzi, M. (2011). Are researchers that collaborate more at the international level top performers? An investigation on the Italian university system. *Journal of Informetrics*, *5*, 204-211. doi: 0.1016/j.joi.2010.11.002

Bordons, M., Aparicio, J., & Costas, R. (2013). Heterogeneity of collaboration and its relationship with research impact in a biomedical field. *Scientometrics*, *96*, 443-466. doi: 10.1007/s11192-012-0890-7

Bordons, M., & Gómez, I. (2000). Collaboration Networks in Science. In B. Cronin & H. B. Atkins (Eds.), *The web of knowledge. A festschrift in honor of Eugene Garfield* (197-213). Medford, NJ: ASIS Monograph.

Costas, R., van Leeuwen, T.N., & Bordons, M. (2010). A bibliometric classificatory approach for the study and assessment of research performance at the individual level : the effects of age on productivity and impact. *Journal of the American Society for Information Science and Technology*, *61*(8), 1564–1581. doi: 10.1002/asi.21348

Cummings, J.N., & Kiesler, S. (2005). Collaborative research across disciplinary and organizational boundaries. *Social Studies of Science*, *35*, 703-722.

Franceschet, M., & Costantini, A. (2010). The effect of scholar collaboration on impact and quality of academic papers. *Journal of Informetrics*, *4*, 540-553. doi: 10.1016/j.joi.2010.06.003

Ganzi, A., Sugimoto, C.R., & Didegah, F. (2012). Mapping world scientific collaboration: authors, institutions, and countries. *Journal of the American Society for Information Science & Technology*, *63*(2), 323-335. doi: 10.1002/asi.21688

Glanzel, W. (2001). National characteristics in international scientific co-authorship relations. *Scientometrics*, *51*(1), 69-115.

He, Z.L. (2009). International collaboration does not have greater epistemic authority. *Journal of the American Society for Information Science and Technology*, 60 (10), 2151–2164. doi: 10.1002/asi.21150

Katz, J.S., & Martin, B.R. (1997). What is research collaboration? *Research Policy*, *26*(1), 1-18.

Kostoff, R.N. (1995). Federal research impact assessment - axioms, approaches, aplications. *Scientometrics*, *34*, 136-145.

Laudel, G. (2002). Collaboration and reward. What do we measure by co-authorships? *Research Evaluation*, *11*(1), 3–15.

Martin, B.R. (1996). The use of multiple indicators in the assessment of basic research. *Scientometrics 36*(3), 343-362.

Martin, B.R., & Irvine, J. (1983). Assessing basic research: some partial indicators for scientific progress in radio astronomy. *Research Policy*, *12*, 61-90.

Melin, G., & Persson, O. (1996). Studying research collaboration using co-authorships. *Scientometrics*, *36*(3), 363–377.

Merton, R.K. (1968). The Matthew effect in science. *Science*, *159*, 56-63.

Moravcsik, M.J. (1984). Life in a multidimensional world. *Scientometrics*, *6*(2), 75-86.

Noyons, E. (2001). Bibliometric mapping of science in a science policy context. *Scientometrics*, *50*(1), 83-98.

Pao, M.L. (1992). Global and local collaborators: a study of scientific collaboration. *Information Processing & Management*, *28*(1), 99-109.

Patel, N. (1973). Collaboration in the professional growth of American Sociology. *Social Science Information*, *12*(6), 77-92.

Persson, O., Glanzel, W., & Danell, R. (2004) Inflationary bibliometric values: The role of scientific collaboration and the need for relative indicators in evaluative studies. *Scientometrics*, *60*(3), 421-432.

Porter, A.L., & Rafols, I. (2009). Is science becoming more interdisciplinary? Measuring and mapping six research fields over time. *Scientometrics*, *81*(3), 719-745.

Reagans, R., & Zuckerman, E.W. (2001). Diversity and productivity: the social capital of corporate R&D teams. *Organization Science*, *12*(4), 502–517.

Seglen, P.O. (1992). The skewness of science. *Journal of the American Society for Information Science*, *43*(9), 628-638.

Tijssen, R.J.W., & de Leeuw, J. (1988). Multivariate data-analysis methods in bibliometric studies of science and technology. *Handbook of Quantitative Studies of Science and Technology, North-Holland, Amsterdam*, 705-740.

Wren, J.D., Kozak, K.Z., Johnson, K.R., Deakyne, S.J., Schilling, L.M., & Dellavalle, R.P. (2007). The write position. A survey of perceived contributions to papers based on byline position and number of authors. *EMBO Reports*, *8*(11), 988–91. doi: 10.1038/sj.embor.7401095

Justificación y estructura de la tesis

Esta tesis doctoral, presentada mediante la modalidad de compendio de artículos, se ha llevado a cabo en el Grupo de Análisis Cuantitativo en Ciencia y Tecnología (ACUTE) del Consejo Superior de Investigaciones Científicas (CSIC); y en colaboración con el Departamento de Estadística de la Universidad de Salamanca. Presenta, por tanto, un marcado carácter interdisciplinar.

La ciencia es uno de los fenómenos de mayor importancia de la sociedad contemporánea, ya que el avance del conocimiento mejora la calidad de vida de las personas y contribuye al desarrollo económico de los países. La colaboración es uno de los rasgos más característicos de la ciencia actual, por lo que estudiar sus causas y efectos ha de ayudar a una mejor comprensión de los mecanismos que gobiernan la ciencia. Esta tesis y los trabajos que la componen surgieron de la preocupación por abordar el estudio de la colaboración desde una perspectiva multivariante, asumiendo que no es viable abordar su estudio como un fenómeno aislado.

La idea que subyace a esta investigación toma coma base el trabajo pionero de Ben Martin (Martin & Irvine, 1983; Martin, 1996) en bibliometría en su concepción de la naturaleza multidimensional de la actividad científica, así como la necesidad de aplicar variables y metodologías diversas para obtener una imagen nítida de la actividad investigadora. Así, se analiza el papel de la colaboración en la ciencia en un sentido amplio y su relación con otras variables, en especial el impacto científico; a distintos niveles de agregación (meso y micro) y haciendo hincapié en las diferencias entre áreas científicas. Para ello, se toma como principal objeto de estudio el CSIC. En este contexto, se presentan aproximaciones metodológicas nuevas o poco utilizadas hasta el momento en bibliometría desde la perspectiva de la estadística multivariante, con especial atención a los métodos Biplot, y del análisis de redes sociales. Las propuestas que se presentan, pretenden ofrecer nuevas posibilidades de análisis y aumentar nuestro conocimiento sobre un fenómeno complejo como el de la colaboración en la ciencia.

La Parte 1 presenta el marco teórico y metodológico, las preguntas de investigación, y las conclusiones generales. En el Capítulo 1, se enmarca el trabajo en el contexto de los estudios cuantitativos sobre ciencia y tecnología, poniendo de manifiesto la relevancia de la bibliometría como herramienta para el estudio del proceso científico. A continuación, se indaga en el surgimiento y relevancia actual de la colaboración en la ciencia, y se resumen los principales esfuerzos y metodologías cuantitativas surgidas para estudiarla. Debido al carácter multidimensionalidad de la actividad científica, se hace hincapié en la necesidad de un análisis integrado de los indicadores para captar el verdadero alcance e influencia de la colaboración en la ciencia. Desde la perspectiva del análisis de datos, se pone de manifiesto el interés de la estadística multivariante en

bibliometría. Por último, se presenta una revisión de las principales técnicas multivariantes aplicadas en el campo.

El Capítulo 2 aborda una serie de aspectos metodológicos fundamentales a todo análisis bibliométrico: la base de datos empleada, el procesamiento y normalización de los datos, varias consideraciones básicas a tener en cuenta, y una clasificación de los indicadores bibliométricos empleados. A continuación, en el Capítulo 3 se exponen la asunciones preliminares sobre las que se ha sustentado este trabajo y se establecen las preguntas de investigación. Finalmente, el Capítulo 4 presenta las conclusiones generales derivadas de esta tesis doctoral, pone en valor el significado de las aportaciones como conjunto y propone futuras líneas de investigación

La Parte 2 está compuesta de cinco capítulos donde se aportan los cuatro artículos publicados o aceptados para su publicación en revistas internacionales. Cada uno de ellos está encaminado a responder a las preguntas de investigación planteadas. El Capítulo 5 presenta un resumen de las publicaciones. Los Capítulos 6 y 7 introducen los métodos Biplot como herramienta de inspección de datos bibliométricos a distintos niveles de análisis (centros e individuos) y se discuten los resultados y patrones encontrados. A través del análisis de redes sociales y los conceptos de centralidad y cohesión, el Capítulo 8 aborda la relación entre el desempeño científico y el rol jugado por los autores en distintas redes de colaboración. El Capítulo 9 explora una nueva fuente de información sobre colaboración, los “agradecimientos”, y propone una metodología para su explotación y tratamiento, combinando minería de texto y Análisis de Correspondencias.

PARTE 1

ANTECEDENTES, METODOLOGÍA Y CONCLUSIONES

1. Introducción

1.1. La medición de la ciencia

La ciencia está compuesta por un cuerpo de conocimientos objetivos y verificables sobre distintas materias. Asimismo, se caracteriza por ser un proceso acumulativo que tiene por objeto la explicación de principios y causas, la formulación de hipótesis o la resolución de problemas a través de la utilización de metodologías adecuadas al objeto de estudio. Por tanto, está sometida a una constante evolución, revisión y contraste crítico.

La ciencia es uno de los fenómenos sociales y culturales más importantes acontecidos en los últimos cinco siglos, especialmente a partir de la II Guerra Mundial. Dado que la ciencia y la tecnología habían sido decisivas para ganar la guerra, los gobiernos de los países más industrializados comenzaron a invertir una inmensa cantidad de recursos en favor del avance científico. Esta época marca el surgimiento de la *big science*, es decir, el desarrollo de grandes proyectos de investigación caracterizados por la colaboración a gran escala y el uso de grandes infraestructuras. Es la época de la bomba atómica, la carrera espacial y, más tarde, la llegada a la Luna. Lógicamente esta gran inversión supuso a su vez un aumento muy significativo en el número de publicaciones científicas, entrando además en escena las editoriales comerciales (Mabe, 2003). Paralelamente, comienzan a surgir órganos gubernamentales para la gestión de la investigación como la *National Science Foundation* en Estados Unidos en 1948. Asimismo, la aparición del Manual de Frascati en 1963, publicado y revisado periódicamente por la OCDE, marca un punto de referencia en la medición y conceptualización de las actividades científicas. Esta vinculación de la ciencia con el desarrollo industrial, además de su profesionalización e incorporación al sistema educativo, hacen que pueda ser entendida como la principal autoridad cognitiva social de nuestro tiempo (Maltrás-Barba, 1996; Sanz-Menéndez & Santesmases, 1996).

La relevancia social de la ciencia, así como su complejo engranaje interno, explican el creciente interés por conocer los distintos procesos y dinámicas que la conforman. Este fenómeno ha sido abordado desde diferentes aproximaciones y perspectivas tales como la sociología, la economía, la política científica, las matemáticas, la filosofía o la historia. Especialmente destacable ha sido el auge experimentado por los estudios cuantitativos de ciencia y tecnología y, más concretamente, la bibliometría bajo la que se han desarrollado nuevos métodos y aproximaciones empíricas para el estudio y comprensión de la ciencia, situándose como un nuevo e interdisciplinar campo de estudio.

Cronológicamente, el nacimiento de la bibliometría o la denominada “ciencia de la ciencia” surge en los años sesenta a raíz de la convergencia de la documentación

científica, la sociología y la historia de la ciencia, con objeto de estudiar la actividad científica como fenómeno y proceso social a través del uso de indicadores y modelos matemáticos (Bordons & Zulueta, 1999). El término fue acuñado por Pritchard (1969) para definir la aplicación de métodos matemáticos y estadísticos a los libros y otros medios de comunicación. No obstante, este campo de estudio, que trata de analizar la ciencia y responder a distintas necesidades (evaluación científica, estudio del proceso científico, vigilancia tecnológica o política científica), ha contado con distintas acepciones a lo largo del tiempo. El término bibliometría es el más usado y extendido. Otra denominación recurrente es cienciometría, término originalmente acuñado por Nalimov como *naukometriya* y empleado en la URSS y el Este de Europa (Egghe, 1988; Wouters, 1999). A pesar de que cienciometría y bibliometría suelen ser utilizados indistintamente, el primero se refiere a un concepto más amplio y no únicamente referido a las los resultados que produce la investigación. Otra denominación más reciente es informetría, de la que Egghe y Rousseau (1990) señalan que se ocupa de la teoría y modelos matemáticos; así como de todos los aspectos relacionados con la información, su almacenamiento y recuperación.

La bibliometría ha estado ligada históricamente a la idea de que es posible representar el conocimiento humano a través de las publicaciones y de los elementos que las componen. Para comprender este concepto ha de entenderse el proceso científico como un sistema de inversiones (*inputs*) y resultados (*outputs*), tal y como propusieron Martin e Irvine (1983) y se ilustra en la Figura 1.1.

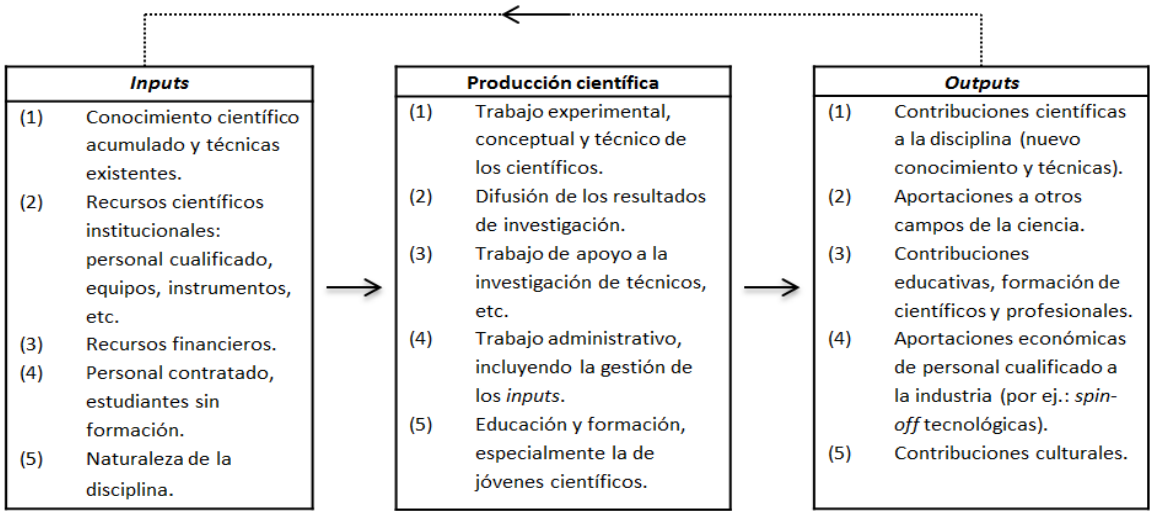


Figura 1.1. Modelo *input-output* propuesto por Martin e Irvine (1983).

Por consiguiente, los *inputs* en un sistema científico están compuestos por una serie de recursos humanos, recursos materiales y conocimientos que a través de un sistema de producción científica generan una serie de *outputs* o resultados. Estos resultados según su naturaleza pueden agruparse en distintas dimensiones (Martin, 1996): dimensión científica (contribuciones al acervo científico), dimensión educacional

(contribuciones a la formación y desarrollo de habilidades), dimensión tecnológica (aportaciones al desarrollo y mejora de las tecnologías) y dimensión cultural (aportes a la sociedad en general). La asunción básica de la bibliometría es que la dimensión más significativa entre los *outputs* que produce el proceso científico es la dimensión científica, es decir, la generación de nuevo conocimiento. Dado que un nuevo conocimiento sólo adquiere valor al ser publicado, las publicaciones científicas son la piedra angular de la ciencia y de los resultados que ésta produce (Kostoff, 1995).

Desde un prisma más social, para la ciencia no basta únicamente con formular nuevos problemas, llevar a cabo experimentos o instaurar la aplicación de nuevos métodos. Comunicar y difundir las investigaciones a través de la publicación de los resultados es básico; determina el carácter acumulativo de la ciencia y permite instaurar un sistema de contribuciones y recompensas. Por un lado, la publicación científica permite asignar una prioridad a los autores de un descubrimiento; por otro lado, los autores de una investigación han de discutir y situar en un contexto apropiado su trabajo, para ello citan trabajos previos que han sido importantes en el desarrollo de su investigación. Esto implica que las citas son un mecanismo de reconocimiento de la influencia del trabajo de otros investigadores (Bornmann & Leydesdorff, 2014). Por tanto, la contribución al progreso y a la generación de nuevo conocimiento científico (entendido como el incremento del conocimiento humano sobre todas las cosas) puede ser medida a través de las publicaciones, que contienen información sobre aspectos tanto cognitivos como sociales del proceso científico.

En la instauración de las publicaciones como medio de difusión e intercambio científico, las revistas han desempeñado un papel central, especialmente las publicaciones a nivel internacional. La aparición de las primeras revistas científicas se remonta al nacimiento de la ciencia moderna a mediados del siglo XVII, concretamente *Philosophical Transactions* y *Journal de Sçavans* comenzaron a publicarse alrededor de 1665 y son consideradas como las primeras revistas científicas. Tal y como apunta van Raan (2004), cada año se añaden al acervo científico de nuestro planeta más de un millón de publicaciones. Este número dividido en múltiples subconjuntos (áreas, países, disciplinas, especialidades), es, en muchos casos, suficiente para permitir análisis cuantitativos que produzcan resultados estadísticamente significativos. Las publicaciones ofrecen una serie de elementos medibles sobre el proceso científico tales como: los nombres de los autores, sus direcciones institucionales, la revista o editorial, las referencias (citas), el año de publicación, la especialidad, las palabras clave o los agradecimientos.

Las primeras bases teóricas sobre las que se sustenta la bibliometría fueron descritas a lo largo del siglo XX en un intento de constatar ciertos patrones inherentes a la actividad científica. Lotka (1926) propuso que, con independencia de la disciplina científica, la distribución de productividad de los autores sigue un modelo general

donde el número de autores que publican *n* documentos es aproximadamente $1/n^2$ de aquellos que publican solamente uno. Esto significa que la gran mayoría de autores publica muy pocos documentos, mientras que unos pocos tienden a concentrar la mayor parte de la producción sobre un tema. Años más tarde, Bradford (1948) formula la ley de dispersión de la literatura científica. La idea subyacente es que la distribución de artículos de una disciplina tiende a concentrarse en un número relativamente pequeño de revistas, mientras que una pequeña proporción de artículos se dispersa en una gran cantidad de revistas. Disponiendo en orden de productividad decreciente las revistas científicas sobre un tema dado es posible establecer un núcleo de revistas específicamente dedicadas a un tema y dos zonas que agrupan aproximadamente el mismo número de artículos, pero donde el número de revistas aumenta según una progresión geométrica del tipo $1: n: n^2$. Además, Price (1976) observó que el crecimiento de la literatura científica es muy superior al de otros procesos sociales y responde a una función exponencial que puede ser expresada como $F(t) = ae^{bxt}$, donde *a* es el tamaño inicial, en el tiempo $t = 0$, y b es la constante que relaciona la velocidad de crecimiento con el tamaño adquirido por la ciencia. No obstante, un crecimiento exponencial hasta el infinito no es factible pues en un momento dado se alcanza un punto de saturación y decrecimiento. Admitido este límite, Price estimó que el crecimiento de la ciencia sigue una curva logística.

Sin embargo, el elemento fundamental que propició el auge de la bibliometría fue el desarrollo de las bases de datos bibliográficas, como respuesta a la gran explosión productiva acontecida tras la II Guerra Mundial. Indudablemente, la creación del *Science Citation Index* a cargo de Eugene Garfield en 1963 constituyó un punto de inflexión al respecto[1]. Su característica más destacada fue incluir las referencias citadas por los autores, permitiendo así encontrar qué documentos citaban a un texto determinado. Durante los años setenta, con objeto de medir las frecuencias de citación de las revistas científicas, Garfield ideó el indicador bibliométrico más conocido, el factor de impacto. En la actualidad, la *Web of Science* (WoS) es la mayor base de datos bibliográfica multidisciplinar a nivel mundial. Esta invención permitió el inicio de análisis estadísticos de la producción científica a gran escala, y supuso el surgimiento de la bibliometría como un campo destacado dentro de los estudios de la ciencia (Wouters, 1999).

Los años setenta y ochenta marcan el despegue y consolidación de la bibliometría. Aquí, sobresalen los trabajos pioneros sobre evaluación científica de Narin (1976), la proposición del análisis de co-citas como modelo para la representación de los campos científicos de Small (1973), la teoría sobre colaboración científica de Beaver y Rosen (1978), el estudio de la interdisciplinariedad (Porter & Chubin, 1985), la necesidad de emplear un conjunto heterogéneo de indicadores (Martin & Irvine, 1983) o el creciente

[1] Inicialmente patrocinado por el *Institute of Scientific Information* (ISI), actualmente es propiedad de la empresa de información *Thomson Reuters*.

interés en el análisis de los parámetros estadísticos de los datos bibliométricos (Schubert & Glänzel, 1983). Ha sido en el ámbito de la evaluación científica donde la bibliometría ha cobrado mayor relevancia debido a factores tales como el constante incremento de los gastos asociados a la investigación, la creciente complejidad y especialización de la ciencia, las dificultades de los comités científicos para definir qué áreas deben priorizarse o la necesidad de una asignación eficiente de los recursos debido a las limitaciones impuestas por el gastos público (Bordons & Zulueta, 1999). Aquí, han emergido los indicadores bibliométricos como metodología para captar y describir diferentes dimensiones de la actividad científica (colaboración, impacto, productividad, interdisciplinariedad, rol en la investigación, etc.) siendo especialmente destacable el uso de las citas como aproximación al concepto de calidad científica (Martin, 1996). En este sentido, hay que destacar la labor llevada a cabo desde los años ochenta por el CWTS de la Universidad de Leiden, en el desarrollo de metodologías e indicadores bibliométricos para la evaluación de la actividad científica.

Por otro lado, la aparición de la revista *Scientometrics* en 1978 supuso un hito importante al dedicarse específicamente a la difusión de los estudios bibliométricos. Esta publicación junto con *Journal of the Association for Information Science and Technology*, *Research Evaluation*, *Research Policy* y, más recientemente, *Journal of Informetrics* constituyen los principales medios de comunicación de las investigaciones bibliométricas. Las décadas siguientes han venido marcadas por un trabajo enfocado al uso aplicado de los indicadores, la importación de metodologías de otras disciplinas, análisis estadísticos cada vez más sofisticados y diversos, así como el uso de nuevas fuentes de datos y posibilidades de análisis. Sin duda, el cambio más notable ha venido producido por el desarrollo de las tecnologías de la información y la comunicación, lo que ha permitido el tratamiento de enormes volúmenes de datos con un menor coste computacional (van Raan, 2004).

De igual modo, la llegada del nuevo milenio ha venido acompañada de nuevas bases de datos y motores de búsqueda, multiplicándose así las posibilidades de obtención de datos. Así, han emergido con especial fuerza *Scopus*, publicada por la empresa *Elsevier*, y *Google Scholar*. En el caso español destacan las bases de datos elaboradoras por el CSIC: *ISOC* relativa a las ciencias sociales y las humanidades, e *ICYT* dedicada a la ciencia y la tecnología. Asimismo, existen numerosas fuentes especializadas en función del área o disciplina de estudio: *PubMed* en biomedicina, *Inspec* en física e ingeniería o *CAS* en química serían algunas de ellas. Además, la masiva generalización de la tecnología ha impulsado la aparición de nuevos canales de comunicación (blogs, redes sociales, etc.), lo que ha favorecido el surgimiento de nuevos indicadores tales como los *altmetrics*. Aunque aún en una fase temprana, similar a la de los indicadores bibliométricos en los años setenta (Bornmann & Leydesdorff, 2014), los datos provenientes de redes sociales como Twitter o Mendeley surgen como información alternativa de la que aún está por discernir su verdadero alcance.

1.2. El fenómeno de la colaboración

Tradicionalmente, los grandes avances científicos han estado asociados a esfuerzos individuales. La historia de la ciencia aclama el rol jugado por genios como Kepler, Newton o Einstein, reflejando la tendencia a vincular grandes ideas con nombres concretos. De forma similar, los premios Nobel reconocen contribuciones científicas sobresalientes de determinados individuos. A pesar del brillo de un selecto número de científicos, lo cierto es que la ciencia no es una tarea individual sino una empresa colaborativa (Merton, 1968). Desde una perspectiva general, el avance científico ha sido posible gracias a la disposición de los científicos a compartir y revelar el conocimiento. A pesar de que la ciencia está sujeta en cada país a las estructuras de poder y las infraestructuras socioeconómicas, el conocimiento científico, idealmente, es supranacional (Subramanyam, 1983).

Más allá de esta consideración general, la colaboración en la ciencia ha adquirido un rol cada vez más relevante, dado que la producción de conocimiento a menudo requiere la participación directa de diversas personas que pueden pertenecer a diferentes instituciones y países (Maltrás-Barba, 1996). Para entender la importancia capital de la colaboración en la ciencia actual es necesario indagar en sus orígenes, para ello debemos regresar al nacimiento de la ciencia moderna en el siglo XVII. Entonces, además de las primeras revistas científicas, apareció en 1665 el primer trabajo científico en colaboración atribuido a Robert Hooke y colaboradores. Desde entonces, la importancia creciente de la ciencia hasta nuestros días ha estado íntimamente ligada a la colaboración científica como uno de sus rasgos más característicos. En la actualidad, la ciencia moderna no puede entenderse, en la mayoría de áreas y disciplinas científicas, sin la colaboración. Beaver y Rosen (1978, 1979a, 1979b) en un trabajo clásico en bibliometría, desarrollaron una exhaustiva teoría de la colaboración científica para explicar los motivos que propiciaron que la ciencia dejara atrás el individualismo en favor de un trabajo colaborativo. Según estos autores, la profesionalización de la ciencia tuvo un papel primordial como impulsor de la colaboración científica.

La profesionalización de la ciencia puede ser entendida como un proceso organizativo dinámico que supuso una reestructuración revolucionaria de lo que, hasta ese momento, había sido un grupo desestructurado de científicos aficionados y a tiempo completo en una comunidad científica (Beaver & Rosen, 1978). Llevó a una redefinición de cómo la ciencia era hecha, quiénes la hacían, quién pagaba por ella o cómo un individuo se convertía en científico. Como modelo de colaboración apuntaron hacia la comunidad científica francesa de principios del siglo XIX. La profesionalización conllevó la institucionalización de la ciencia a través de la fundación de sociedades científicas, centros de investigación, revistas científicas y; finalmente, la aceptación social de la ciencia. Tanto las relaciones internas (entre científicos) como las externas

(con la sociedad) llevaron al establecimiento de reglas comunes y obligaciones; al mismo tiempo que se definían las relaciones con las personas que en un futuro iban a formar parte de la comunidad científica. La institucionalización de la ciencia condujo a la creación de una jerarquía profesional. El grupo más profesionalizado era la élite, que controlaba la financiación, el derecho a decidir el trabajo científico y el acceso a la propia élite. En consecuencia, para aquellos científicos que pertenecían a la élite suponía una forma de controlar las nuevas ideas y conocimientos que provenían de investigadores más jóvenes. En cambio, para los científicos jóvenes, la colaboración con la élite les permitía maximizar sus *outputs*, publicar en revistas más prestigiosas y una mayor disponibilidad de recursos.

En términos de magnitud, durante el siglo XIX el trabajo en colaboración, medido a través de publicaciones en co-autoría, experimentó un crecimiento paulatino, pasando de un 2% en 1800 a aproximadamente un 7% en 1900 (Beaver & Rosen, 1979b). No fue hasta después de la Segunda Guerra Mundial y el surgimiento de la *big science* que la colaboración se convirtió en una característica distintiva de la ciencia moderna, especialmente a partir de finales de los años sesenta (Bordons & Gómez, 2000). Aunque la mayor parte de la colaboración en la ciencia no puede ser clasificada bajo el término *big science*, lo cierto es que este tipo de investigación muestra algunas de las características más reconocibles de la colaboración actual tales como la interdisciplinariedad, la dependencia de financiación económica o la necesidad de llegar a acuerdos sobre la propiedad intelectual de los resultados.

Un ejemplo de la explosión colaborativa acontecida a partir de la segunda mitad del siglo XX puede hallarse en Wren et al. (2007) en el ámbito de la medicina. Los autores muestran como en 1966 el 40% de los trabajos publicados en medicina eran firmados por un único autor; mientras que la participación de un elevado número de autores era muy escasa. En las décadas siguientes esta tendencia comienza a invertirse rápidamente. Así, en 2006 el patrón de colaboración en medicina ha sufrido un giro radical y más del 90% de los trabajos han sido publicados en colaboración.

Otra muestra del rol predominante de la colaboración en la ciencia actual puede encontrarse en Gazni, Sugimoto y Didegah (2012), quienes analizaron catorce millones de trabajos en colaboración y observaron como las tasas de colaboración han pasado de un 69% en el año 2000 a un 78% en 2009. De igual forma, el número medio de autores por documento ha crecido de forma progresiva, pasando de 3,3 en el año 2000 a 4,1 en 2009. Este incremento en el número de investigadores involucrados en un trabajo científico se entiende como un reflejo de la creciente dificultad y complejidad de la investigación (Wang, Wu, & Pan, 2014).

El predominio de la colaboración en la ciencia actual viene explicado por una serie de factores, algunos inherentes a la actividad investigadora y otros externos a ella (Cummings & Kiesler, 2005; Sonnenwald, 2007):

- ✓ *Científicos:* considerando la creciente especialización de los científicos y la interdisciplinariedad de muchos temas de investigación, cada vez es más necesario combinar diferentes enfoques para afrontar muchos de los problemas que se plantean en ciencia, como por ejemplo, los relativos al medio ambiente. Estas aproximaciones interdisciplinares pueden ser enriquecedoras y en algunos casos conducir a la aparición de nuevas especialidades científicas y/o nuevas disciplinas en las fronteras de la ciencia. No obstante, la colaboración, y en especial aquélla que se desarrolla entre investigadores con distinta especialización, con frecuencia se asocia a un desarrollo más lento de la investigación por la necesidad de coordinación y discusión de los resultados.

- ✓ *Económicos*: debido a los elevados costes de la investigación, la colaboración permite compartir recursos y abaratar gastos. Colaborar permite el acceso a instrumentos o instalaciones de coste muy elevado que de otra forma serían muy difíciles de obtener. Hoy día, las colaboraciones entre autores o instituciones muy alejadas geográficamente son totalmente factibles gracias al desarrollo y difusión de las tecnologías de la información y la comunicación.

- ✓ *Políticos*: a nivel internacional las agencias financiadoras fomentan la colaboración, sobre todo la internacional, como forma de estimular la unidad política y el mutuo entendimiento entre países. En este contexto, la colaboración entre países cuyas relaciones oficiales son tensas puede funcionar como un catalizador para la paz, por ejemplo, a través de investigaciones que buscan solucionar problemas comunes entre las regiones. Al mismo tiempo, puede resultar desafiante pues las políticas nacionales de cada país pueden no apoyar tales esfuerzos.

- ✓ *Personales y sociales*: existen un amplio espectro de factores sociales que pueden proporcionar la base para iniciar una colaboración tales como la búsqueda de nuevas ideas, la necesidad de interacción personal, la formación investigadora y tutoría de jóvenes investigadores o la curiosidad intelectual.

Además, desde un punto de vista más pragmático la colaboración puede resultar de interés para los investigadores al favorecer (Melin, 2000; Beaver, 2001): la productividad (el reparto de tareas puede permitir a un científico trabajar en varios proyectos simultáneamente), un aumento de la calidad (el trabajo en equipo aumenta las sinergias y distintas personas pueden aportar diferentes conocimientos a una investigación), la visibilidad (publicación en revistas más prestigiosas), un incremento de la competitividad (afrontar proyectos más ambiciosos) o la reducción de riesgos (participar en varios proyectos aumenta las probabilidades de éxito).

Desde la perspectiva de la política científica, no es de extrañar que la investigación colaborativa sea fomentada por parte de las agencias financiadoras debido a la

limitada disponibilidad de recursos y a la fuerte dependencia de la financiación para poder llevar a cabo las investigaciones. A modo de ejemplo, se pueden citar algunas iniciativas de la *National Science Foundation*[2] o de la Comisión Europea[3]. A nivel europeo, se exige la colaboración entre equipos de distintos países miembros de la UE en los proyectos de investigación del programa marco y en las redes de excelencia para favorecer la cohesión dentro de la Unión; y se promueve la movilidad del personal investigador para favorecer los vínculos internacionales. Las políticas nacionales también tratan de favorecer la integración de los grupos nacionales en la ciencia internacional promoviendo la colaboración más allá de las propias fronteras. En nuestro país, tanto el Plan Nacional como los Planes propios de las diversas comunidades autónomas tienen entre sus objetivos promover la colaboración a distintos niveles: entre individuos, fomentando la creación de grupos de investigación; entre disciplinas, promoviendo la interdisciplinariedad de la investigación; entre sectores institucionales, favoreciendo los contactos entre universidad y sector privado; y entre centros, mediante la creación de redes -especialmente en biomedicina-.

Teniendo en cuenta lo expuesto anteriormente, resulta evidente que la colaboración es un fenómeno complejo. Existen en la literatura diferentes aproximaciones y metodologías para su estudio desde ópticas muy diversas que oscilan desde los métodos cuantitativos basados en *outputs*, principalmente la bibliometría, a otros desde la sociología o la filosofía de carácter más teórico y centradas en el estudio de la colaboración como proceso (Gónzalez-Alcaide & Gómez Ferri, 2014). Tradicionalmente, los estudios bibliométricos han abordado la colaboración en la ciencia a través del análisis de co-autoría, es decir, la firma conjunta de varios investigadores en las publicaciones científicas. A través de esta información se han desarrollado numerosos indicadores que miden diferentes aspectos de la colaboración. En este contexto, *Web of Science* ha sido ampliamente utilizada al incluir información exhaustiva sobre autores y direcciones (Bordons & Gómez, 2000). Las ventajas de analizar la colaboración a través de indicadores basados en co-autoría son que los resultados son verificables y reproducibles, y que es posible analizar grandes conjuntos de datos.

No obstante, el uso de la co-autoría como aproximación para el estudio de la colaboración en la ciencia presenta algunas limitaciones. Por un lado, tal y como señalan Laudel (2002) y Melin y Persson (1996), equiparar co-autoría a colaboración presenta el riesgo de pasar por alto algunas actividades colaborativas que no quedan reflejadas en la firma conjunta de publicaciones. Por otro lado, bajo el prisma de la co-autoría se da por hecho que todos los autores firmantes de un trabajo han participado en el desarrollo del mismo, pero puede suceder que algunas autorías sean gratuitas u honoríficas (Cronin, 2001). A pesar de estas limitaciones, está ampliamente aceptado

[2] www.nsf.gov/od/oia/programs/stc
[3] http://ec.europa.eu/research/iscp/index.cfm?lg=en&pg=fp7

que la co-autoría proporciona información de interés acerca de la estructura y dinámicas de colaboración entre investigadores. Distintos autores han señalado que las co-autorías gratuitas u honoríficas y la no inclusión de autores que han realizado contribuciones relevantes suelen ser más propias de la colaboración intramural (entre miembros de un mismo grupo, departamento o instituto) que de la extramural (entre diferentes instituciones) (Glänzel & Schubert, 2004). Por esta razón, cuando se considera la colaboración llevada a cabo entre investigadores afiliados a distintas instituciones estas limitaciones presentan menor incidencia.

Llegados a este punto se hace necesario esclarecer el concepto de colaboración. La colaboración puede definirse como una forma intensa de interacción que tiene lugar en un contexto social entre dos o más investigadores, permitiendo compartir esfuerzos, tareas, hallazgos y habilidades con objeto de lograr una meta común, generalmente, la producción de nuevo conocimiento (Melin & Person, 1996; Sonnenwald, 2007). Este objetivo de producir conocimiento generalmente deriva en la elaboración de publicaciones científicas, aunque también puede dirigirse hacia el desarrollo tecnológico, de patentes o de *software* (Bozeman, Fay, & Salde, 2013).

Aquí se nos plantea una cuestión interesante ¿cuál debe ser el grado de interacción mínimo entre dos investigadores que trabajan juntos para considerarlo como colaboración? A pesar de que no es sencillo alcanzar una conclusión sin fisuras, Katz y Martin (1997) señalan que la colaboración debe incluir a investigadores que trabajen juntos en un proyecto o durante una parte importante de la investigación. Esta acepción es coherente con la definición de "autor" que propone el Comité Internacional de Revistas Médicas (ICMJE, 2015) en cuyas recomendaciones establecen que un investigador para considerarse autor de un trabajo científico ha de haber participado en las siguientes tareas: 1) concepción o diseño del trabajo; o adquisición, análisis o interpretación de los datos de la investigación; y 2) elaboración del borrador o revisión crítica importante del contenido intelectual; y 3) aprobación final de la versión del trabajo para publicar. En este sentido, las aportaciones de los colaboradores que no cumplan estos requisitos deberían ser reconocidas en la sección de los agradecimientos (Claxton, 2005).

En este escenario, los agradecimientos que aparecen en las publicaciones científicas surgen como un indicador complementario sobre las interacciones que tienen lugar durante el proceso colaborativo. De hecho, su utilidad como fuente de información sobre sub-autoría científica ya fue señalada décadas atrás por autores como Heffner (1981). Definiéndolos formalmente, son un acto voluntario de reconocimiento que aparece a través de un código implícito en la conducta profesional de los investigadores siendo, generalmente, expresiones de gratitud relativas a diversos tipos de apoyo recibidos en la investigación (Cronin, 1995). A diferencia de las citas, que son reconocimientos formales de deuda intelectual, los agradecimientos presentan un

significado más informal, personal y singular (Giles & Council, 2004). No obstante, esta fuente de información ha sido escasamente analizada en la literatura porque sólo recientemente ha comenzado a ser incluida en alguna base de datos.

También basados en la co-autoría, otro tipo de aproximación interesante para estudiar la colaboración es la basada en análisis de redes sociales. El análisis de redes sociales se basa fundamentalmente en la teoría de grafos, rama de las matemáticas que tiene por objeto el estudio de los vínculos o relaciones existentes entre pares de objetos, de forma que es posible representar y estudiar funciones matemáticas. Su uso en bibliometría permite indagar en las relaciones e influencias entre científicos, los flujos de conocimiento o la estructura y evolución de los campos científicos. Vinculado principalmente al análisis de citas desde los años setenta, los trabajos de Barabási et al. (2002) y Newman (2004) han extendido su aplicación a las redes de co-autoría. No obstante, aún queda por explotar su verdadero potencial como herramienta para el análisis de la colaboración en la ciencia. Un enfoque particularmente interesante en análisis de redes sociales es el basado en el concepto de centralidad (Freeman, 1979), que permite discernir entre los diversos roles desempeñados por los nodos de una red.

En este contexto, resulta interesante recordar el concepto de "colegios invisibles"[4] (Price & Beaver, 1966), entendido como el círculo social que se crea entre investigadores a través de congresos o estancias de investigación en otros centros. Más recientemente, Zuccala (2006) los define como un círculo de científicos que comparten intereses de investigación, publican sobre la materia y se comunican formal e informalmente, pudiendo pertenecer a instituciones geográficamente dispersas. Estos colegios constituyen un mecanismo poderoso capaz de controlar el prestigio de un científico o el destino de las nuevas ideas. Crane (1972) asimiló y desarrolló este concepto en el sentido de que los integrantes de un círculo social únicamente conocen a algunos integrantes del mismo, pero están influenciados por otras personas a las que no están conectadas directamente. El análisis de redes sociales puede constituir una herramienta interesante para profundizar en estas relaciones.

Por otro lado, cabe señalar en este punto el surgimiento de los términos "visualización" y "mapeo" de la ciencia como etiquetas habituales en la literatura bibliométrica para hacer referencia a metodologías muy diversas en cuanto a su complejidad y objetivos. Así nos encontramos englobadas bajo esta acepción a metodologías relativas al propio análisis de redes sociales, a técnicas de geolocalización o a métodos de análisis multivariante (Noyons, 2001; Börner, Chen, & Boyack, 2003); metodologías que, en mayor o menor grado, han sido empleadas para abordar el estudio de la colaboración en la ciencia.

[4] El término "colegios invisibles" tiene su origen en las relaciones informales y comunicaciones entre los científicos y pensadores de Reino Unido a mediados del siglo XVII y que, posteriormente, dieron lugar a la creación de la *Royal Society* de Londres.

1.3. La multidimensionalidad de la ciencia

En el primer apartado se explicó el modelo de *inputs* – *outputs* relativo al proceso científico y se subrayó que la bibliometría se centra, principalmente, en la dimensión científica al considerarse que las publicaciones son el medio primordial para la difusión de nuevo conocimiento. Tal y como señalaba Martin (1996), la dimensión científica encierra a su vez una naturaleza multidimensional. En consecuencia, se han desarrollado multitud de indicadores en un esfuerzo tanto por captar distintos aspectos relacionados con el desempeño científico, como debido a los múltiples procedimientos de cálculo (van Leeuwen, Visser, Moed, Nederhof, & van Raan, 2003). Aunque desde la aparición del índice-h y derivados (Hirsch, 2005) ha existido cierta tendencia a reducir diferentes dimensiones del desempeño científico a un solo indicador bibliométrico (Wildgaard, Schneider, & Larsen, 2014), la creencia general dentro de la comunidad bibliométrica es que uno o dos indicadores no son suficientes para describir la actividad científica de un país, un área o disciplina o un investigador. Esto se traduce en que la ciencia ha de ser entendida como un fenómeno social en el que todos los elementos interactúan y donde no es posible aislar los efectos de unos factores sobre otros; ni que éstos sean descritos sin ambigüedades. Por tanto, las aproximaciones uni o bidimensionales no son realistas en este entorno (Moravsick, 1984).

El estudio de la colaboración como parte del engranaje de la ciencia requiere una gran diversidad de métodos e indicadores, tanto por las distintas formas en que la colaboración se manifiesta, como por las distintas posibilidades de monitorizarla a través de las publicaciones científicas. Asimismo, para medir el verdadero alcance de la colaboración es necesario tener en cuenta las interacciones y los efectos que produce sobre diferentes aspectos de la actividad científica. De acuerdo con la literatura podemos establecer tres dimensiones sobre las que, principalmente, se ha analizado el alcance y los efectos que produce la colaboración:

- ✓ *Producción y actividad científica*: la actividad científica de, por ejemplo, un grupo de investigación está influenciada por factores intrínsecos a la colaboración como el número de investigadores participantes, la disponibilidad de personal de apoyo o el acceso a equipos e instalaciones.

  En este sentido, se ha descrito en la literatura una fuerte relación entre colaboración y productividad científica, que contribuye a explicar la alta productividad de la élite científica dentro de cada disciplina (ver por ej. Pao, 1992). No obstante, la asociación entre estas variables depende de las áreas (Abramo, D'Angelo, & Di Costa, 2009), por un lado debido a la diferente naturaleza de la investigación en unas y otras, pero también porque la colaboración tiene sus costes asociados, como son los derivados de la necesidad de coordinar la investigación

(Cummings & Kiesler, 2005), que deben ser menores que sus beneficios para que la colaboración sea "rentable" a los investigadores. Asimismo, esta asociación es más clara para los países desarrollados que para los que están en vías de desarrollo, donde una realidad más compleja puede socavar los esfuerzos colaborativos (Duque et al., 2005; Tovainen & Ponomariov, 2011).

Otros posibles factores que influyen sobre la relación entre colaboración y productividad son la edad de los investigadores, su género, categoría profesional, entorno de trabajo y financiación (Lee & Bozeman, 2005; Carayol & Matt, 2004; Bozeman & Gaughan, 2011; Abramo, D'Angelo, & Di Costa, 2011). Dada la influencia de la colaboración sobre los resultados de la actividad científica, es importante introducir variables asociadas a la misma en los estudios sobre la actividad de los investigadores a nivel individual (Costas, van Leeuwen, & Bordons, 2010).

- ✓ *Impacto científico:* como la calidad (mérito intrínseco de la investigación) e importancia (influencia potencial) de una investigación no puede ser medida de forma directa, las citas recibidas por un trabajo científico (influencia real) son empleadas como un indicador indirecto de calidad para estimar el impacto de una contribución sobre el avance del conocimiento (Martin & Irvine, 1983).

  En lo que se refiere a los efectos positivos de la colaboración sobre la calidad de la investigación, se ha observado un mayor impacto de las publicaciones realizadas en colaboración entre varios autores frente a las realizadas en solitario (Persson, Glänzel, & Danell, 2004), así como un incremento del impacto -medido a través de citas- al aumentar el alcance de la colaboración (nacional, internacional) (Katz & Hicks, 1997; Abramo, D'Angelo, & Solazzi, 2011). Los beneficios de la colaboración internacional parecen ser superiores a los de la colaboración nacional, especialmente para los países periféricos, aunque con variaciones según disciplinas y países (Glänzel, 2001; Jarneving, 2010; Franceschet & Costantini, 2010). De igual modo, se ha descrito que los autores muy productivos tienden a colaborar más habitualmente con otros investigadores que comparten sus mismos temas de investigación, pero citan tanto a autores de sus mismos temas como a otros de distintas especialidades (Ding, 2011).

- ✓ *Interdisciplinariedad*: la ciencia es cada vez más una actividad interdisciplinar, especialmente entre especialidades próximas (Porter & Rafols, 2009). Esto se traduce en que los equipos de investigación tiene una composición más heterogénea, lo que posibilita que cada integrante aporte conocimientos e ideas propias, que se genere una gama más amplia de objetivos y que los resultados sean también más diversos (Jha & Welch, 2010). En la literatura se ha puesto de manifiesto que la colaboración entre distintas instituciones involucra un mayor

grado de heterogeneidad e interdisciplinariedad, además de mayores beneficios en términos de impacto, que la colaboración intra-mural (Franceschet & Constantini, 2010; Bordons, Aparicio, & Costas, 2013). Este mayor impacto de la investigación se atribuye a que la diversidad de puntos de vista estimula la creatividad e innovación (Reagans & Zuckerman, 2001).

1.4. ¿Por qué una aproximación multivariante?

Desde la perspectiva del análisis de datos, la multidimensionalidad de la ciencia expuesta en el apartado anterior pone de manifiesto la necesidad de abordar el estudio de la ciencia a través de un conjunto heterogéneo de indicadores bibliométricos. En este contexto, con multitud de variables y factores relacionados, es donde emerge la utilidad de la estadística multivariante para delinear y describir la compleja estructura subyacente de los datos, y las interacciones que tienen lugar entre los distintos indicadores bibliométricos.

El análisis multivariante es la parte de la estadística que estudia, analiza, representa e interpreta una serie de variables estadísticas que han sido medidas sobre una muestra de *n* individuos, es decir, es el conjunto de métodos destinados a describir e interpretar los datos provenientes de la observación de varias variables (Cuadras, 2014). Cuando se lleva a cabo un análisis multivariante generalmente se dispone de una matriz de datos, aunque la información de entrada también puede estar contenida en matrices de distancias o similaridades que miden el grado de discrepancia entre las observaciones.

En su forma más simple, nos referimos a una matriz rectangular con n filas y p columnas, donde sobre n elementos de la población se han medido p variables. En el ámbito bibliométrico, una matriz típica puede ser la que se describe a continuación. Supongamos que sobre *n* investigadores $(r_1, \dots, r_n)$, que trabajan en una determinada disciplina, se han medido p indicadores bibliométricos $(x_1, \dots, x_p)$ relativos a su desempeño científico. Sea $X_{ij}(r_{ij})$ la observación del indicador x_j sobre el investigador r_i. De forma que se dispone de una matriz de datos X con $n\ x\ p$ dimensiones[5].

$$X = \begin{bmatrix} x_{11} & x_{12} & \dots x_{1p} \\ x_{i1} & x_{22} & \dots x_{2p} \\ \vdots & \vdots & \vdots \\ x_{ni} & x_{n2} & \dots x_{np} \end{bmatrix} = \begin{bmatrix} x'_1 \\ x'_2 \\ \vdots \\ x'_n \end{bmatrix} = \left[x_1, x_2 \cdots x_p\right]$$

La estructura interna de la matriz de datos con respecto a las observaciones o unidades de muestreo (tipo de relación o dependencia entre las variables), la tipología y distribución de los datos, así como el propio objetivo del análisis que se pretende llevar a cabo, son los factores que van a determinar la selección de la técnica multivariante a aplicar. Tal y como señalan Egghe y Rousseau (1990), en análisis multivariante se dan cita dos aspectos relevantes. El primero de ellos es la descripción

[5] A esta estructura estándar de dos vías también es posible incorporar una tercera vía, como pueden ser distintos períodos temporales o distintos conjuntos de observaciones (siguiendo el ejemplo, investigadores).

de una variable como una función de un conjunto de otras variables, es decir, involucra múltiples variables que están correlacionadas entre sí. Por tanto, el carácter multivariante de los datos no descansa únicamente en el número de variables sino en las múltiples combinaciones posibles entre las mismas. Un segundo aspecto hace referencia a la reducción de la dimensión. La idea subyacente es que aprovechando la estructura de relaciones entre las variables es posible simplificar el problema en estudio a unas pocas dimensiones (generalmente dos) con mínima pérdida de información. Geométricamente, técnicas como el Análisis de Componentes Principales o el Análisis de Correspondencias se pueden conceptualizar como diagramas de dispersión en un espacio p dimensional, donde las dimensiones (ejes) están definidos o explican la información contenida por las variables de interés.

Más allá de estas consideraciones generales, los datos bibliométricos poseen una serie de características que dificultan la aplicación de test estadísticos, y que sin embargo, resultan muy apropiadas para la aplicación de técnicas de análisis multivariante.

- ✓ *Variables correlacionadas*: además de la mencionada relación entre colaboración científica y otras dimensiones tales como impacto, productividad o interdisciplinariedad; los estudios bibliométricos combinan con asiduidad distintas medidas relativas a una misma dimensión para describir y caracterizar la actividad científica desde diferentes ópticas. Estos dos aspectos propician que, generalmente, las variables empleadas estén bastante correlacionadas. Desde la perspectiva de la estadística uni y bivariante la relación entre las variables objeto de estudio puede resultar en problemas de colinealidad, como en el caso de modelos regresión múltiple, cuando algunas de las variables explicativas están correlacionadas entre sí. Sin embargo, las técnicas multivariantes permiten estudiar la estructura de correlaciones subyacente a la matriz de datos y reducir el problema a un número menor de dimensiones con mínima pérdida de información.

- ✓ *Distribuciones muy asimétricas*: la distribución de distintas variables, como por ejemplo las citas, ha sido un tema ampliamente debatido en la comunidad bibliométrica debido a su carácter asimétrico. Aunque no existe un acuerdo en la literatura, se considera que dada una distribución de publicaciones con *n* autores, la colaboración entre número reducido de científicos se ajusta a una distribución de Poisson o a una binomial negativa, mientras que la investigación ligada a la *big science* parece responder a una ley de potencia $1/n^k$ con k del orden de 4 a 6 (Beaver, 1986). Gupta, Kumar, y Rousseau (1998) investigan distintas distribuciones y concluyen que una distribución de Poisson truncada se ajusta a los campos donde predomina la publicación en solitario, mientras que una distribución geométrica sirve para describir el número de autores.

En cuanto a la distribución de las citas, Price (1965) sugirió originalmente su asimetría que posteriormente fue puesta de manifiesto por Seglen (1992) en un influyente trabajo. Seglen ilustró la asimetría de las citas a través de varios ejemplos a nivel de artículo y de revista, y teniendo en cuenta distintas ventanas de citación. Para una muestra aleatoria de 938 artículos, obtenida del *Science Citation Index* en el período 1985-1989, observó que cerca del 60% de los artículos no habían recibido ninguna cita pasados tres años de su publicación, mientras que una proporción cada vez menor de artículos tendía a concentrar un mayor número de citas. De igual modo, si la perspectiva es desde el ámbito de las revistas, el rasgo característico de las distribuciones de citas es la asimetría. En concreto, Seglen (1992) analizó la contribución acumulada de citas de tres revistas del área de bioquímica, distribuidas en 20 percentiles conteniendo cada uno un 5% del número total de artículos que integraban cada revista. Halló que un 15% de los artículos de las revistas aglutinaban un 50% de las citas que habían recibido las publicaciones, mientras que un 50% de los artículos concentraban casi el 90% de las citas recibidas por la revista.

Investigaciones más recientes han ahondado en diferentes características de las citas tales como poder ser representadas a través de leyes de potencias, y la plausibilidad de este fenómeno a distintos niveles de agregación (Egghe, 2005; Albarrán, Crespo, Ortuño, & Ruiz-Castillo, 2011). Recientemente, una investigación llevada a cabo por Ruiz-Castillo y Costas (2014) sugiere que a pesar las grandes diferencias en términos de productividad, las distribuciones de los datos son similares entre las disciplinas. Este hallazgo pone de manifiesto que a pesar de que las disciplinas difieren ampliamente en términos de producción y citas, una explicación de las variaciones intra-campo podría ser suficiente para describir este fenómeno.

- ✓ *Menor potencia de contraste de los test estadísticos*: la asimetría de la ciencia se traduce en una gran variabilidad entre las observaciones. De igual modo, el tamaño de las muestras desempeña un papel relevante. A nivel macro se suele trabajar con muestras de gran tamaño compuestas por millones de documentos, mientras que a nivel micro las muestras pueden llegar a ser muy pequeñas. Tanto la gran variabilidad como la disparidad en los tamaños muestrales propicia una menor potencia de contraste de los test estadísticos. En cambio, desde una perspectiva multivariante mayor variabilidad de los datos significa más información, y diferentes tamaños muestrales no tienen por qué minimizar la validez de resultados en los casos de técnicas multivariantes donde no son necesarios ciertos supuestos de partida (normalidad, homocedasticidad, linealidad).

1.5. Análisis multivariante en bibliometría

El auge de la bibliometría en las últimas décadas ha propiciado un uso cada vez mayor de técnicas estadísticas avanzadas, así como la aparición de infinidad de métodos de visualización de la información. Como se ha puesto de manifiesto en apartados anteriores, la multidimensionalidad y asimetría de la ciencia, entre otras características, hacen de las técnicas de análisis multivariante una herramienta muy a tener en cuenta en el ámbito bibliométrico al permitir, además de una visualización de los datos, estudiar la compleja estructura subyacente de los datos y las relaciones entre las variables.

Como medida orientativa del grado de integración de las técnicas multivariantes en la comunidad bibliométrica, se ha llevado a cabo una búsqueda bibliográfica en la base de datos *Web of Science*. Se ha tomado como referencia la revista *Scientometrics* por ser la publicación con mayor tradición en el campo. Esta publicación, frente a otras revistas como *Journal of the Association for Information Science and Technology* o *Research Evaluation*, que aceptan una amplia gama de temas y aproximaciones, está dedicada específicamente al estudio de la ciencia y la tecnología desde una perspectiva cuantitativa. Aunque en los últimos años *Journal of Informetrics* se ha situado como la publicación más enfocada en aspectos matemáticos y estadísticos, comenzó a publicarse recientemente, concretamente, en el año 2007.

En el diseño de la sentencia de búsqueda[6] se han incluido distintas variantes y acrónimos de las técnicas de análisis multivariantes más clásicas y/o extendidas. En términos de recuperación es complicado obtener unos resultados óptimos, puesto que algunas técnicas pueden haber sido empleadas en la investigación, pero no haber sido referenciadas en el título, resumen o palabras clave del documento. Para mejorar la exhaustividad, se han introducido algunos términos o metodologías tradicionalmente aplicadas en los estudios bibliométricos como *co-citation analysis* y *co-word analysis*,

[6] *SO= Scientometrics AND (TS="MULTIVARIATE ANALYSIS" OR TS="MULTIVARIATE ANALYSES" OR TS="MULTIVARIATE STATISTICS" OR TS="MULTIVARIATE STATISTICAL ANALYSIS" OR TS="MULTIVARIATE STATISTICAL ANALYSES" OR TS="MULTIVARIATE METHODS" OR TS="MULTIVARIATE MODELS" OR TS="MULTIVARIATE TECHNIQUES" OR TS= "PRINCIPAL COMPONENT ANALYSIS" OR TS= "PCA" OR TS= "MDS" OR TS= "MULTIDIMENSIONAL SCALING" OR TS= "CLUSTER ANALYSIS" OR TS="CLUSTERING" OR TS= "CORRESPONDENCE ANALYSIS" OR TS= "CORRESPONDENCE FACTOR ANALYSIS" OR TS="CFA" OR TS="MULTIPLE CORRESPONDENCE ANALYSIS" OR TS="MCA" OR TS= "FACTOR ANALYSIS" OR TS="EFA" OR TS="CFA" OR TS="STRUCTURAL EQUATION MODELING" OR TS="SEM" OR TS= "CANONICAL ANALYSIS" OR TS="CANONICAL CORRELATION ANALYSIS" OR TS="CCA" OR TS= "PRINCIPAL COORDINATES ANALYSIS" OR TS= "PCoA" OR TS= "MANOVA" OR TS="DISCRIMINANT ANALYSIS" OR TS="CO-CITATION ANALYSIS" OR TS="COCITATION ANALYSIS" OR TS="CO-WORD ANALYSIS" OR TS="COWORD ANALYSIS" OR TS="LOG-LINEAR" OR TS="LOG LINEAR" OR TS="VOSVIEWER") Período de tiempo=1978-2014. Tipos de documento: (ARTICLE).*
-> SO = Nombre de publicación; TS = Tema.

donde se trabaja con matrices de co-ocurrencia y en las que se suelen emplear como métodos de representación técnicas multivariantes. En cambio, se ha optado por no considerar el término mapeo (*mapping*), pues aunque en ocasiones se emplea para referirse a técnicas de análisis multivariante, también engloba a numerosas técnicas de visualización de la información. Por otro lado, el desarrollo de *software* específico ha estimulado la difusión y aplicación de técnicas estadísticas cada vez más avanzadas. Estos desarrollos han estado principalmente basados en teoría de grafos, con potentes aplicaciones como *CiteSpace* (Chen, 2006) o *Pajek* (Batagelj & Mrvar, 2013). No obstante, también han surgido programas basados en métodos multivariantes como *VOSviewer* (van Eck & Waltman, 2010) que emplea una metodología para la visualización de similaridades (VOS) que, bajo ciertas condiciones (ver van Eck & Waltman, 2007), es equivalente a un Escalamiento Multidimensional o *Multidimensional Scaling* (MDS) no métrico de Sammon (1969). La introducción de estos términos ha permitido mayor exhaustividad en la recuperación de la información, sin que la precisión de los mismos pueda verse afectada de forma significativa.

La Figura 1.2 muestra los resultados de la búsqueda agrupados en distintos períodos temporales. Se observa como en los primeros años, que van desde la aparición de la revista en 1978 hasta finales de los años ochenta, la presencia de artículos publicados en *Scientometrics* con aplicación de técnicas multivariante es mínima. Hay que tener en cuenta que en este período, la bibliometría comienza a establecerse como campo de estudio y muchas de las bases conceptuales y metodológicas se asentaron durante aquellos años. Por lo tanto, a pesar de esta baja incidencia es posible encontrar algunos trabajos clave al respecto. De hecho, Tijssen y de Leeuw (1988) ya señalaron la incipiente importancia del análisis multivariante en bibliometría, poniendo además de manifiesto la necesidad de un conocimiento pormenorizado de las distintas técnicas.

Los años noventa y la primera década del nuevo milenio suponen la incorporación y asentamiento de las técnicas multivariantes entre las metodologías de análisis de datos de los investigadores en bibliometría. El último lustro ha significado la explosión definitiva en cuanto su aplicación en el campo. El crecimiento de la comunidad bibliométrica, el desarrollo de software especializado o el tratamiento masivo de información han sido factores determinantes para entender la vulgarización acontecida en los últimos años.

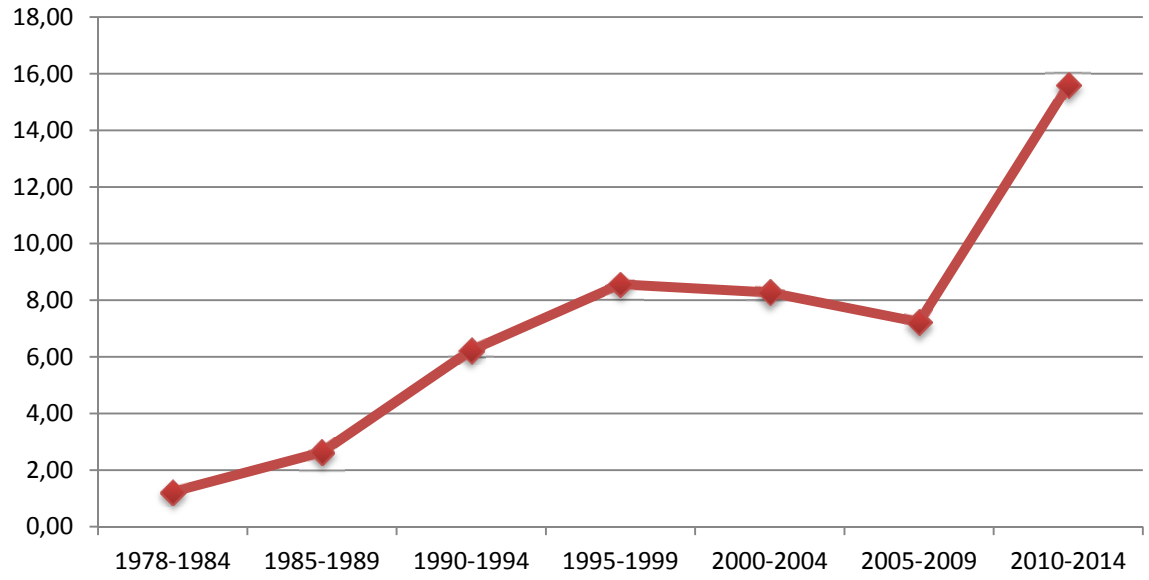


Figura 1.2. Evolución temporal del porcentaje de artículos que aplican métodos multivariantes en la revista *Scientometrics.*

Entre las aplicaciones incipientes más destacadas pueden citarse los trabajos de Small y colaboradores (Small & Sweeney, 1985; Small, Sweeney, & Greenlee, 1985), quienes utilizaban Análisis de Cluster y MDS para la creación de mapas de la ciencia. Para ello, usaban el *Science Citation Index* y análisis de co-citación. Las citas eran previamente normalizadas mediante el índice de Jaccard y como algoritmo aglomerativo empleaban el vecino más próximo. Finalmente, para revelar la estructura de especialidades y los flujos de co-citación entre los distintos conglomerados aplicaban un MDS.

Otro ejemplo puede encontrarse en McCain (1983), quien emplea un MDS no métrico y Análisis de Cluster para estudiar la estructura de co-citación entre 42 autores prominentes en el ámbito de la macroeconomía entre 1972 y 1981. Los conglomerados resultantes se correspondían con diferentes escuelas de pensamiento, identificándose a su vez dos dimensiones principales: una formada por autores orientados hacia aspectos cuantitativos y modelos matemáticos, y otra relacionada con la edad y la formación.

Dentro del Análisis de Cluster es posible identificar dos tipos fundamentales de métodos de clasificación: jerárquicos y no jerárquicos. En los primeros, la clasificación resultante tiene un número creciente de clases anidadas mientras que en el segundo las clases no son anidadas. Las diferencias entre los distintos algoritmos residen en la forma de definir la distancia entre una observación y un grupo que contenga varias observaciones, o entre dos grupos de observaciones. En la literatura bibliométrica puede destacarse el uso de los métodos jerárquicos, en concreto, el método de Ward (Ward, 1963) debido a que suele producir conglomerados homogéneos y con tamaños similares. Este método no calcula distancias entre conglomerados, sino que forma los clusters de forma que se maximice la homogeneidad intra-cluster. Los conglomerados

obtenidos en cada paso, son los que minimizan la suma de los cuadrados intra-grupos. Citando ejemplos recientes, este método ha sido aplicado para comparar los resultados obtenidos mediante distintas formas de medir las frecuencias de co-citación entre autores (Jarvening, 2008); así como en el sistema universitario español en el área de informática para clasificar a las universidades, y a su personal académico, según diversos aspectos relacionados con el desempeño científico (Ibáñez, Larrañaga, & Bielza, 2013).

También pueden hallarse estudios que emplean Análisis de Componentes Principales (PCA), técnica cuyo objetivo es la reducción de la dimensión de los datos con el fin de simplificar el problema en estudio (ver entre otros Hotelling, 1933). Consiste en encontrar transformaciones ortogonales de las variables originales para conseguir un nuevo conjunto de variables incorreladas (Componentes Principales), que se obtienen en orden decreciente de importancia y de las que se espera sean capaces de recoger la mayor parte de la variabilidad presente en los datos. Entre otras aplicaciones, esta técnica se ha utilizado para proyectar clusters de documentos agrupados por tema mediante el uso de palabras clave (Polanco, François, & Kleim, 1998), así como para estudiar las similitudes entre un conjunto de indicadores a nivel individual (Todeschini, 2011). Otra técnica muy utilizada en los estudios bibliométricos es el Análisis Factorial, siendo generalmente empleado como método de estimación de la matriz factorial las Componentes Principales. El Análisis Factorial tiene por objeto explicar las relaciones entre un conjunto de variables observables, a través de un número reducido de variables hipotéticas o latentes (factores) que se obtienen a partir de las correlaciones entre las variables originales. Ejemplos al respecto pueden encontrarse en Ramesh-Babu y Singh (1998), quienes detectaron once factores latentes que puedan afectar a la producción científica de los investigadores. Trabajos más recientes lo han aplicado al análisis de co-citas entre autores (Schneider, Larsen, & Ingwersen, 2009) o para indagar en la capacidad de los *rankings* universitarios para medir diferentes constructos (Safón, 2013).

Durante los años ochenta también aparecen las primeras aplicaciones de técnicas de representación simultánea de datos multidimensionales, siendo el Análisis de Correspondencias (Benzécri, 1973) la técnica seleccionada por los investigadores. Desde la perspectiva del análisis de datos, resulta interesante la propuesta de Tijssen, de Leeuw y van Raan (1987). Estos autores proponen la utilización de un Cuasi Análisis de Correspondencias (de Leeuw & van der Heijden, 1985) para lograr un mejor ajuste en el caso de matrices de datos con ceros estructurales o datos faltantes, situación muy habitual en el ámbito bibliométrico cuando se manejan datos de citas. Sin embargo, esta técnica comenzó a extenderse en el campo a raíz de una serie de estudios posteriores llevados a cabo en el CNRS. Así, encontramos una primera publicación que explora las posibilidades del Análisis de Correspondencias para revelar la estructura de colaboración internacional entre países y campos científicos (Okubo,

Miquel, Frigoletto, & Dore, 1992). En concreto, los autores emplean una matriz compuesta por 98 países y 8 áreas científicas y encuentran que los países pueden ser principalmente clasificados en dos clusters: por un lado, países con una elevada actividad en colaboración internacional en física y, por otro lado, países que principalmente colaboran en biología y medicina clínica.

Siguiendo esta línea, un estudio posterior de Miquel, Ojasoo, Okubo, Paul y Doré (1995) profundiza en los patrones de colaboración internacional entre los 48 países con mayor producción, en una selección de disciplinas científicas, durante el período 1981-1992. Asimismo, publicaciones posteriores de este grupo emplean la técnica en el análisis jerárquico de la co-autoría en las redes de colaboración (Abd el Kader, Ojasoo, Miquel, Okubo, & Doré, 1998), y en el análisis de patentes (Doré, Dutheuil, & Miquel, 2000). El Análisis de Correspondencias también ha sido aplicado por Nagpaul (1995) para evaluar la contribución de los investigadores de las universidades de la India a las revistas de mayor impacto internacional; y por Pereira y Escuder (1999) para identificar las especialidades con mejor rendimiento científico en la comunidad científica brasileña de ciencias de la salud.

Aunque en menor medida, pueden hallarse en *Scientometrics* la aplicación de otras técnicas multivariantes centradas en el estudio de las diferencias entre grupos tales como el Análisis Discriminante o el Análisis Multivariante de la Varianza (MANOVA). El Análisis Discriminante permite identificar qué combinación lineal de variables discrimina mejor entre dos o más grupos, es decir, busca las direcciones de máxima separación entre los grupos. Una vez encontrada esa combinación, la función discriminante puede ser utilizada para clasificar nuevas observaciones. Zhou, Guo, Ho y Wu (2007) lo utilizan para estudiar los patrones de publicación en geo-estadística y construir funciones discriminantes que validen la clasificación previamente obtenida mediante Análisis de Cluster. En cambio, si el objeto es la caracterización de las diferencias entre los vectores de medias a través de contrastes estadísticos, la técnica apropiada es un MANOVA. Este método es empleado por Park y Kang (2009) para analizar las diferencias en los flujos de conocimiento científico y tecnológico en el ámbito coreano por campo científico. Además, para indagar en la combinación de variables que mejor separa a los grupos aplican un Análisis Discriminante, encontrando que el número de artículos y patentes citadas son las variables con mayor poder discriminante.

Por último, entre otros ejemplos, el *software VOSviewer* ha sido empleado para revelar los principales temas abordados en las reseñas editoriales publicadas por las revistas *Nature* y *Science* (Waaijer, van Bochove, & van Eck, 2011); y para mostrar los flujos de citación entre patentes a través de la *International Patent Classification* (Leydesdorff, Kushnir, & Rafols, 2014).

2. Conceptos metodológicos

En esta sección se describen cuatro elementos fundamentales en todo estudio bibliométrico: la fuente de datos empleada; el procesamiento y normalización de los datos; algunas consideraciones metodológicas básicas; así como una clasificación y descripción general de los indicadores empleados.

2.1. La fuente de datos: *Web of Science*

En la actualidad existe una variada colección de bases de datos, tanto multidisciplinares como especializadas, lo que permite abordar el estudio de cualquier área científica. Emerge aquí la necesidad de seleccionar la fuente adecuada para el área objeto de estudio. Las bases de datos difieren en cuanto a su cobertura temática, los criterios de selección de las revistas y/o documentos, así como en sesgos geográficos o lingüísticos. Estas características deben ser tenidas en cuenta previamente para seleccionar la fuente de información más adecuada (Bordons & Zulueta, 1999).

Para la realización de este trabajo se han empleado las bases de datos bibliográficas de la plataforma *Web of Science* publicada por la empresa *Thomson Reuters*. Su carácter multidisciplinar, la exhaustiva inclusión de autores, instituciones y de las referencias citadas por los autores son sus características más destacadas. Su explotación permite proporcionar una visión de la producción científica de un país en su vertiente más internacional y obtener indicadores de actividad, impacto o colaboración, entre otros. Ha sido, y aún es en la actualidad, la principal herramienta empleada en los estudios bibliométricos.

A continuación, se describen las bases de datos multidisciplinares que integran *Web of Science* y se discuten sus principales características y limitaciones[7] (Archambault, Vignola-Gagne, Coté, Lariviere, & Gingras, 2006):

Bases de datos multidisciplinares:

- ✓ *Science Citation Index Expanded* (SCI-Expanded): recoge la literatura internacional publicada desde 1900 en revistas científicas relativas a las ciencias experimentales. Incluye 8.693 revistas y más de 150 disciplinas. Semanalmente añade unos 19.000 registros y 423.000 referencias
- ✓ *Social Sciences Citation Index* (SSCI): ofrece una cobertura de las revistas de ciencias sociales desde 1956 en adelante. Indiza 3.168 revistas de 55 disciplinas e

[7] http://ip-science.thomsonreuters.com/mjl/

incluye documentos de otras 3.500 revistas. Aproximadamente se añaden 2.900 registros y 60.000 referencias citadas por semana.

- ✓ *Arts & Humanities Citation Index (A&HCI)*: base de datos especializada en arte y humanidades con una cobertura desde 1975 hasta la actualidad. Se incluyen más de 1.747 revistas en 27 disciplinas, así como una selección de documentos de otras 6.800 revistas. Se incorporan 2.300 registros y 15.250 referencias citadas por semana aproximadamente.

Las tres bases de datos mencionadas junto al *Conference Proceedings Citation Index* (CPCI), el *Book Citation Index* (BCI), el *Current Chemical Reactions* (CCR) y el *Index Chemicus* (IC) conforman la denominada *Web of Science Core Collection*[8] que recoge más 12.000 de revistas, 160.000 actas de congresos y más de 60.000 libros.

Características:

- ✓ *Multidisciplinariedad*: cubre un gran abanico de especialidades recogiendo los elementos publicados en la ciencia más internacional o *mainstream science*. En la actualidad incluye más de 12.000 revistas entre ciencias experimentales, ciencias sociales, arte y humanidades.
- ✓ *Selectividad*: la selección de revistas se realiza teniendo en cuenta criterios de calidad científica, calidad formal y reconocimiento por parte de la comunidad investigadora (citas).
- ✓ *Exhaustividad*: recoge la información completa de las revistas seleccionadas y todos los aspectos medibles sobre las publicaciones (autores, direcciones, referencias y citas, etc.).

Limitaciones:

- ✓ *Sesgo lingüístico*: la cobertura tiende a favorecer a los países anglosajones así como a las revistas publicadas en inglés, siendo Estados Unidos y Reino Unido los más países más favorecidos.
- ✓ *Sesgo documental*: la producción científica que recoge es fundamentalmente la publicada en revistas científicas.
- ✓ *Sesgo disciplinar*: ofrece una mejor cobertura de las disciplinas básicas en detrimento de las más aplicadas y con mayor orientación local, por lo que en general no ofrece una cobertura óptima de la dimensión nacional.
- ✓ *Normalización*: a pesar de los esfuerzos, *Web of* Science presenta importantes inconsistencias en el procesamiento de las direcciones institucionales y los nombres de autores.

[8] http://thomsonreuters.com/content/dam/openweb/documents/pdf/scholarly-scientific-research/fact-sheet/web-of-science-core-collection.pdf

En función de estas características y limitaciones, la cobertura que ofrece *Web of Science* es, en general, bastante superior para las ciencias experimentales que para las ciencias sociales o humanidades (Moed, 2005). La Tabla 2.1 muestra una estimación de la cobertura de la base de datos por disciplinas en el año 2002. Por un lado, las ciencias experimentales más básicas y que presentan una mayor orientación internacional son las mejor recogidas (> 80%). Por otro lado, las más aplicadas, como las ingenierías, presentan una menor cobertura (40-60%) debido a que suelen emplearse como canales de comunicación actas de congresos o *proceedings papers*. Entre las ciencias sociales, aquellas más próximas a las ciencias de la salud son las mejor recogidas (60-80%). Para humanidades y artes la base de datos ofrece una cobertura moderada (< 40%), debido a que suelen emplearse más a menudo otras tipologías como libros y capítulos de libro.

Tabla 2.1. Cobertura disciplinar de *Web of Science* (Moed, 2005).

Excelente (> 80%)	Muy buena (60-80%)	Buena (40-60%)	Moderada (< 40%)
Biología molecular y bioquímica	Física y química aplicada	Matemáticas	Otras ciencias sociales
Ciencias biológicas afines a los humanos	Ciencias biológicas afines a animales y plantas	Economía	Humanidades
Química	Psicología y psiquiatría	Ingeniería	Artes
Medicina clínica	Geociencias		
Física y astronomía	Otras ciencias sociales afines con la medicina y la salud		

Conscientes de las limitaciones mencionadas, *Thomson Reuters* ha tratado de paliar y globalizar durante los últimos años la base de datos. Así, con la creación en 2008 de los *Conference Proceedings Citation Index Science and Social Sciences* y, en 2010, el *Book Citation Index* se ha mejorado la cobertura para aquellas disciplinas donde los artículos en revistas no son el principal canal de comunicación. Asimismo, hemos asistido a la incorporación de un significativo número de revistas regionales, 1.600 entre 2007-2009. En el *Arts & Humanities Citation Index* durante el período 2005-2010 se indexaron 434 nuevas revistas, lo que supuso un aumento cercano al 40% con respecto al año 2005 (Testa, 2011).

A pesar de los sesgos analizados, la selección de revistas recogida en las distintas bases de datos de *Web of Science* responde al principio, ya señalado, como ley de Bradford o ley de dispersión de la literatura científica (Bradford, 1948). Esto se traduce en que no es necesario disponer de toda la bibliografía sobre un tema, sino que un pequeño número de revistas concentra la mayor parte de los trabajos relevantes sobre el mismo. Por tanto, únicamente las revistas que reúnan una serie de requisitos y

criterios pueden ser seleccionadas para su inclusión en la base de datos[9]. Estos criterios incluyen el cumplimiento de una serie de estándares de publicación tales como periodicidad, corrección formal o revisión por pares. Asimismo, la publicación ha de reunir unos requisitos mínimos en cuanto a su cobertura temática y visibilidad, es decir, en cuanto a su presencia en bases de datos, audiencia y citas recibidas y, sobre todo, relevancia del equipo editorial.

En lo que respecta a la explotación de la base de datos como fuente de información sobre colaboración científica, *Web of Science* ha ido muy utilizada al incluir a todos los autores firmantes de una publicación, así como los lugares de trabajos de los autores. Además, desde agosto de 2008 ha comenzado a incluir la información de la sección de los agradecimientos que, generalmente, aparece en la parte final de los trabajos, previamente a las referencias[10]. Esta información se incluye en un campo llamado *Funding Acknowledgment* (FA) que a su vez está divido en tres subcampos: *Funding Agency (*FO) que contiene la información de las agencias u organismos financiadores; G*rant Number* (FG) que proporciona, si lo hubiese, el código de identificación del proyecto; y *Funding Text* (FT) que contiene el texto completo de la sección de los agradecimientos tal y como aparece originalmente en la publicación. En la Tabla 2.2 se muestra un ejemplo de un registro extraído de esta base de datos.

Tabla 2.2. Ejemplo de un registro del campo *funding acknowledgment* de WoS.

Funding Agency	Grant Number	Funding Text
Spanish Ministerio de Ciencia e Innovación-FEDER. Aragon regional government (Geotransfer research group)	CGL2009-13390	The research has been financed by project CGL2009-13390 of the Spanish Ministerio de Ciencia e Innovación-FEDER, as well as by the Aragon regional government (Geotransfer research group). Our research benefited from the data and comments by H. Perea about geomorphic indices. T. Salvador and F.J. Gracia Abadias collaborated in the geophysical survey of the Ametler graben infill and the numerical treatment of morphometric data, respectively. OSL dating was made by Laboratorio de Datación y Radioquímica of the Universidad Autónoma de Madrid.

Queda patente en el ejemplo que los agradecimientos contienen información, entre otras cosas, sobre otro tipo de interacciones que tienen lugar durante el proceso colaborativo y que va más allá de los datos sobre co-autoría. Previamente, esta información no había estado disponible en las bases de datos bibliográficas, por lo que su extracción y explotación a gran escala no era viable. Por tanto, la inclusión de esta información abre nuevas vías para el análisis de la colaboración en la ciencia. No obstante, su explotación resulta complicada ya que se trata de información no estructurada, es decir, texto en lenguaje natural.

[9] http://wokinfo.com/essays/journal-selection-process/
[10] http://wokinfo.com/products_tools/multidisciplinary/webofscience/fundingsearch/

2.2. Tratamiento y normalización de los datos

Todas las aportaciones que se presentan en esta tesis doctoral han empleado la información incluida en las siguientes base de datos de *Web of Science*: *Science Citation Index Expanded* (SCIE), *Social Sciences Citation Index* (SSCI) y el *Arts & Humanities Citation Index* (A&HCI), para analizar la actividad científica de áreas, centros y autores. Para la asignación de crédito a los trabajos realizados en colaboración se ha empleado un conteo completo o *full counting*, de forma que a todos los autores y/o centros se les ha asignado el mismo peso en la colaboración, independientemente del número de firmantes o su posición en el pie de autor (*byline)*.

En primer lugar, previamente al tratamiento y normalización de los datos, es necesario establecer la clasificación o delimitación temática empleada. En las diferentes publicaciones se ha empleado la clasificación de revistas en subcampos o disciplinas científicas (disciplinas WoS) elaborada por *Thomson Reuters*. Cada revista puede aparecer clasificada hasta en seis disciplinas diferentes, respondiendo así a la creciente tendencia a publicar en revistas de campos afines (Lewison, 1999). Para agrupar las disciplinas a nivel meso, los subcampos se clasificaron en diez grandes áreas partiendo de criterios similares a los del *Current Contents* (ver Anexo I). En el caso de los trabajos centrados en la actividad científica del CSIC, la producción se asigna según las ocho grandes áreas científico-técnicas de la institución de forma que la producción es asignada a centros y los centros a áreas (Bordons et al., 2014).

Como apuntan de Lange y Glänzel (1997), la actividad científica de cualquier ente viene determinada por la información sobre autores y direcciones contenida en el documento. Tal y como se señaló en la sección anterior, los datos referentes a instituciones y autores de las publicaciones presentan importantes inconsistencias y errores. Por tanto, un elemento esencial para que los resultados de cualquier análisis bibliométrico sean fiables es el tratamiento y normalización de estos datos. Para las instituciones, el principal problema reside en las diferentes variantes de firma para una única institución; problema que se multiplica debido al uso del inglés o las lenguas maternas. Asimismo, puede suceder que una institución cambie de nombre a lo largo del tiempo. En cuanto a los autores, las inconsistencias son debidas, por un lado, a que la base de datos emplea algoritmos para estructuras de nombres anglosajones; por otro lado, muchos autores emplean varias variantes de firma lo que complica enormemente la normalización. Además, se dan problemas de homonimia debido a apellidos y nombres muy comunes.

Para el tratamiento y normalización de los datos de autores e instituciones en esta investigación se ha seguido una metodología similar a la empleada por el grupo ACUTE en otros trabajos (ver por ejemplo, Zulueta, Cabrero, & Bordons, 1999; Gómez et al.,

2009; Bordons, et al. 2014). Los datos son tratados de forma semiautomática a través de una serie de programas desarrollados 'ad hoc' que combinan bases de datos relacionales y minería de datos con objeto de normalizar los nombres de autores e instituciones. El modelo relacional, aunque sometido a múltiples mejoras y actualizaciones, se basa en la estructura original propuesta por Fernández, Cabrero, Zulueta y Gómez (1993). Su núcleo está compuesto por la relación entre la información del propio documento, sus diferentes direcciones (campo WoS *Address*), autores (campo WoS *Author*) y tablas maestras (códigos de normalización, direcciones ya codificadas) que contienen información que facilitará la automatización del proceso.

A nivel meso, la producción científica es asignada a instituciones, lo que implica la normalización de la información recogida en el campo *Address*. La identificación y normalización de las direcciones se lleva a cabo de forma semi-automática a través de una aplicación informática. A continuación se resumen sus fases más importantes (Morillo, Santabárbara, & Aparicio, 2013):

- ✓ Generar una base de datos maestra a partir de las instituciones con direcciones codificadas manualmente en estudios previos, y que cuentan con un código asignado que las identifica de forma unívoca con un centro.

- ✓ Para las nuevas direcciones que no están codificadas, un algoritmo genera un listado de centros candidatos para codificar automáticamente tantas direcciones como sea posible. Para ello, se emplea un programa para el procesamiento y normalización de las direcciones a través de consultas. Este programa estudia cada dirección comparándola con la información de la base de datos maestra, identificando posibles códigos candidatos a cada uno de los cuales les asigna un peso en función de su similitud con la dirección a codificar. Por último, teniendo en cuenta los pesos asignados, y en función de unos umbrales de verificación, codifica la dirección o le asigna posibles candidatos.

- ✓ Facilitar la codificación manual del resto de las direcciones sin candidatos o cuyo listado de centros candidatos no haya superado los umbrales para la codificación automática.

Para el estudio a nivel micro de los autores y redes de colaboración es necesario el tratamiento y normalización de la información que recoge *Web of Science* en el campo *Author*. Para llevar a cabo este proceso, la aplicación desarrollada combina la información relativa a las direcciones, centros, autores colaboradores y firmas colaboradoras. La normalización de los autores se lleva a cabo de la siguiente forma:

- ✓ La aplicación inicia un proceso de asignación de centros a autores, que ayuda a la identificación normalizada de autores, puesto que saber si una firma pertenece a un centro determinado es relevante de cara a compararla con otra firma parecida. Este proceso permite discriminar entre autores con el mismo nombre pero que tienen diferente lugar de trabajo. Para los documentos con una única dirección, se vinculan todos los autores a dicha dirección. Para los documentos con dos o más direcciones, se descartan centros y autores ya identificados para comprobar si al final del proceso queda únicamente una dirección por clasificar que se pueda asignar al autor restante.

- ✓ El procesamiento de los autores se lleva a cabo a través de un algoritmo de similitud que, ante dos firmas similares, compara centros y autores colaboradores para decidir si se trata de la misma persona. Este algoritmo tiene en cuenta los apellidos más comunes reduciendo el peso y, por consiguiente, la medida de similitud entre firmas. Asimismo, ignora las palabras vacías.

- ✓ Finalmente, la aplicación propone posibles candidatos de nombre normalizado para los autores, siendo la asignación final validada manualmente.

2.3. Niveles de análisis y consideraciones generales

A la hora de abordar cualquier análisis bibliométrico, resulta imprescindible considerar una serie de aspectos metodológicos para que los resultados obtenidos sean consistentes y fiables. En este apartado, se exponen los diferentes ámbitos de aplicación de los indicadores bibliométricos, sus limitaciones inherentes, así como varias asunciones básicas relativas a su aplicación para el estudio de la colaboración en la ciencia.

El análisis de la actividad científica se puede caracterizar tanto en función de los indicadores empleados para describirla, como por el tamaño de las diferentes unidades a los que estos indicadores van a ser aplicados. Por tanto, la validez y los enfoques van a variar en función de si el objeto de estudio es la producción científica de un país, de un área, de una disciplina o de un grupo de investigación (Costas, 2008). Por esta razón, se han descrito en la literatura diferentes niveles de agregación: macro, meso y micro (Vinkler, 1988). Así, el mayor nivel de agregación sería el análisis de la actividad científica a nivel mundial, mientras que la unidad de estudio más pequeña serían los investigadores a nivel individual (Tabla 2.3).

Tabla 2.3. Unidades objeto de estudio en bibliometría (Vinkler, 1988).

Niveles de análisis	Organización	Temática	Publicación
Macro	- Mundo - País/es - Sector institucional	- Ciencia en general - Área científica	- Una base de datos
Meso	- Centro/s - Departamento/s - Instituto/s	- Disciplina	- Colección de publicaciones
Micro	- Grupo de investigación - Autores	- Proyecto	- Una revista - Un artículo

A continuación se describen los diferentes niveles de agregación mencionados (Costas, 2008):

- ✓ *Nivel macro*: en este nivel se incluyen los estudios que abordan el análisis de grandes unidades tales como países, áreas científicas o la producción científica mundial. Debido al gran tamaño de los conjuntos de datos con las que se trabaja, los errores debidos a datos faltantes o problemas en el proceso de normalización presentan una influencia pequeña sobre el conjunto. En consecuencia, los indicadores bibliométricos presentan gran validez y fiabilidad a este nivel. Así, se hallan en la literatura numerosos estudios a nivel macro que analizan la producción

científica a nivel mundial o por países (ver por ejemplo Luukkonen, Persson, & Siversten, 1992; Glänzel, 2000; Guerrero-Bote, Olmeda-Gómez, & Moya-Anegón, 2013).

- ✓ *Nivel meso*: este nivel engloba a todos aquellos estudios que tienen por objeto describir la actividad científica de centros de investigación, universidades o disciplinas científicas. Este tipo de análisis son los más habituales en la literatura bibliométrica, debido a que los subconjuntos que se analizan son lo suficientemente grandes como para que los errores por problemas de normalización no presenten gran incidencia y no suelen requerir un tratamiento masivo de datos. Además, responden en muchas ocasiones a necesidades de evaluación o sirven como herramientas de apoyo a la política científica. Algunos ejemplos de estudios publicados en la literatura a este nivel pueden ser la descripción de los patrones de colaboración en la universidad de Harvard (Gazni & Didegah, 2011), el análisis de la investigación sobre cáncer en Rusia (Lewison & Markusova, 2010) o el estudio de la actividad científica del CSIC (González-Albo, Moreno, Morillo, & Bordons 2012).

- ✓ *Nivel micro*: se trata del menor nivel de agregación. Aquí se encuadran los estudios bibliométricos que abordan el análisis de grupos de investigación, autores individuales, proyectos o revistas. Aunque sujeto a ciertas limitaciones, este nivel resulta de sumo interés dado que el desempeño científico individual tiene una importancia crucial para el funcionamiento de la empresa científica. A este nivel el volumen de documentos a analizar puede ser pequeño, lo que puede redundar en mayores sesgos y errores, así como en una menor validez de los métodos estadísticos tradicionales (Glänzel & Wouters, 2013). Por ello, los análisis son extremadamente sensibles a errores y se requiere de un proceso de normalización de datos muy exhaustivo. Además, debido a que la influencia de factores no científicos es especialmente importante en los estudios sobre la producción de investigadores, emerge aquí con mayor intensidad la necesidad de apoyar los resultados con métodos cualitativos e incorporar variables de contexto o personales tales como la edad, el sexo, la categoría profesional, la movilidad o la docencia. La utilidad de los indicadores bibliométricos en los procesos de evaluación de la investigación ha sido el factor determinante para explicar el creciente interés y demanda de los estudios bibliométricos a nivel micro. Esto es debido a que incrementan la objetividad en la toma de decisiones, por lo que cada vez son más requeridos por responsables de política científica, por gestores y por los propios científicos; no obstante, se incide en la necesidad de complementarlos con juicios de expertos (Costas, van Leuween, & Bordons, 2010).

Independientemente del nivel de análisis, hay que atender una serie de consideraciones básicas relativas a la aplicación e interpretación de los indicadores bibliométricos (Bordons & Zulueta, 1999; Moed 2000). Por un lado, no deben efectuarse comparaciones en bruto entre distintas áreas o disciplinas debido a que los hábitos de publicación y actividad difieren entre sí. Factores tales como el tamaño del campo (número de investigadores, documentos y revistas), la naturaleza básica o aplicada de la investigación que se lleva a cabo, el grado de desarrollo de la comunidad científica o la necesidad o no de apoyo económico pueden desempeñar un papel relevante sobre el nivel de actividad científica. Estos factores son interdependientes y sus efectos pueden solaparse y cambiar a lo largo del tiempo (Vinkler, 2010).

Estas diferencias son especialmente notables entre áreas (ciencias experimentales, ciencias naturales, ingenierías, ciencias sociales y humanidades); aunque los hábitos de producción también varían en función de las disciplinas que la integran. De igual modo, los canales de comunicación comúnmente empleados para difundir las investigaciones son diferentes entre las áreas. Así, en las ciencias experimentales prima la publicación a través de artículos, en las ingenierías las actas de congresos o *proceedings papers* son una tipología muy habitual, mientras que libros y capítulos de libros constituyen los canales principales en humanidades.

Las diferencias entre áreas han sido especialmente descritas en el caso de las citas (ver por ejemplo, Moed, 2005). En campos con una alta productividad y un rápido envejecimiento de la literatura (por ej.: biología celular), el número medio de citas recibidas por publicación es generalmente mucho más elevado que en disciplinas menos productivas y de menor densidad de citación (por ej.: matemáticas) (Waltman & van Eck, 2013). Por tanto, resulta ineludible emplear indicadores normalizados para llevar a cabo comparaciones entre conjuntos de datos de distintas áreas. Para la normalización de las citas es habitual emplear como referencia las citas relativas al mundo de las disciplinas de publicación de los artículos. La clasificación de revistas en categorías temáticas realizada por *Web of Science* es una de las más utilizada con este fin (ver Anexo I).

En lo que respecta al estudio de la colaboración en la ciencia, estas afirmaciones y consideraciones son igualmente aplicables. A pesar de que la colaboración desempeñe un papel preponderante en la ciencia actual existen grandes diferencias por áreas (Cronin, 2001). En las ciencias experimentales la colaboración es habitualmente un requisito imprescindible a causa de la fuerte dependencia de la financiación, y de grandes equipos procedentes de diferentes instituciones y/o países. Además, la colaboración internacional, con la participación de un número elevado de centros, suele ser característica de la investigación en disciplinas como astronomía, física de partículas o genética. En cambio, en otras áreas menos dependientes de la

financiación, con una orientación más teórica y/o de interés más regional o local puede suceder que la colaboración no sea un requisito imprescindible o que los colaboradores más pertinentes pertenezcan a la misma institución o región. Ejemplos de este tipo de investigación son humanidades o una parte de las ciencias sociales. No obstante, la colaboración también ha crecido en estas áreas, especialmente en los últimos años (Gazni et al., 2012).

En términos de co-autoría, la distinta naturaleza y características de las áreas científicas se traduce en grandes diferencias en el número medio de centros firmantes, tipo de colaboración predominante o número medio de autores por documento. Asimismo, las áreas y disciplinas tienen sus propias reglas en lo que respecta a los hábitos de co-autoría (Birnhtolz, 2006). De hecho, la importancia de algunas contribuciones varía en función el área, mientras en humanidades y ciencias sociales la redacción y elaboración del documento es un elemento crucial, en las ciencias naturales y biomédicas el análisis de los datos juega un papel muy importante (Haustein & Larivière, 2015). De igual modo, la posición de firma no sigue un patrón común en todos los campos, sino que el orden viene determinado según las costumbres y prácticas en el país y área de publicación (Liu & Fang, 2014). La primera y última posición están consideradas como las más relevantes, especialmente en las ciencias experimentales, asumiéndose que el primer autor es el principal responsable del trabajo experimental, mientras que el último autor puede ser entendido como el supervisor y fuerza motriz de la investigación (Tscharntke, Hochberg, Rand, Resh, & Kruass, 2007). No obstante, hay que tener en cuenta que en campos como matemáticas o economía, la firma por orden alfabético es una costumbre dentro de la disciplina, aunque lo cierto es que se trata de una tendencia en declive (Waltman, 2012).

En definitiva, estos elementos ponen de manifiesto que la actividad científica no presenta patrones uniformes. Por ello, el estudio de la colaboración en la ciencia debe ser contextualizado en función del país, el área y la disciplina de estudio. De igual modo, los indicadores empleados deben haber sido formulados y calculados para responder a la naturaleza del estudio, y no presentar inconsistencias que distorsionen los resultados.

2.4. Clasificación de los indicadores bibliométricos

Los indicadores bibliométricos son medidas que permiten describir diferentes aspectos cuantitativos de la actividad científica y que pueden atribuirse a una o varias unidades de estudio. Cualquier indicador, por definición, debería representar de manera fidedigna los patrones o tendencias que pretende medir o pronosticar. Los principales requisitos que un indicador debe reunir son: adecuación estadística, fiabilidad y eficacia en la medición, buena cobertura, sensibilidad a los cambios, difusión y aceptación, homogeneidad en la medición temporal, eficacia y facilidad de interpretación (Moore & Shiskin, 1967). Por otro lado, debido a los diferentes patrones de actividad científica entre áreas y disciplinas expuestos en la sección anterior, resulta necesario incorporar métodos de normalización para que las comparaciones entre investigadores, centros o instituciones relativas a distintos dominios científicos sean pertinentes.

En base a estas consideraciones, los indicadores bibliométricos se pueden clasificar de acuerdo a las metodologías de normalización empleadas. Además, en función de los aspectos medidos por los indicadores, éstos pueden ser agrupados en distintas dimensiones (de colaboración, de impacto, de producción, etc.). Debido a la distinta naturaleza y objetivos de ambas clasificaciones, su exposición conjunta ofrece una visión integrada sobre las distintas facetas de los indicadores bibliométricos.

2.4.1. Indicadores según la aplicación de estándares de referencia

En función de las unidades o sistemas analizados (investigadores, equipos, disciplinas, organizaciones, países, etc.), y la normalización y/o ponderación de los datos es posible clasificar de forma general los indicadores bibliométricos de la siguiente manera (Vinkler, 1988; Vinkler 2010):

- ✓ **Indicadores en bruto o *gross indicators***: son aquellos que miden un único aspecto de la actividad científica, para un solo conjunto y que presentan un único nivel jerárquico. Así, los indicadores en bruto puede ser definidos de la siguiente forma:

$$GI = \sum_{i=1}^{N} w_e \, . \, e_i$$

Donde e_i es el elemento *i*-ésimo del conjunto, N es el número total de elementos y w_e es un posible factor de ponderación. Ejemplos de este tipo de indicadores son: el número de publicaciones de una institución durante un período de tiempo dado, el número de investigadores adscritos a un departamento, el número de citas que han recibido los trabajos publicados por un grupo de investigación determinado o

el número de individuos con los que un investigador ha colaborado. Su aplicación a, por ejemplo, la actividad científica de una organización a lo largo del tiempo puede producir interesantes resultados sobre su crecimiento, estancamiento o decrecimiento. En cambio, no son apropiados para comparar entre unidades de distinto tamaño o naturaleza.

- ✓ **Indicadores complejos o *complex indicators***: son aquellos que hacen referencia a varios conjuntos o a un único conjunto con varios niveles jerárquicos. Los indicadores complejos caracterizan un aspecto bibliométrico concreto y tienen un significado bien definido. Pueden definirse de forma general como:

$$CI = A \cdot f \cdot B$$

Donde f representa una operación matemática, A y B son el número de elementos del conjunto [A] y [B] respectivamente. Los indicadores compuestos pueden estar normalizados, los indicadores en bruto no. La normalización entre dos conjuntos puede calcularse como:

$$CI = \frac{w_a \cdot A}{w_b \cdot B} = \frac{\sum_{i=1}^{A} w_a \cdot a_i}{\sum_{i=1}^{B} w_b \cdot b_i}$$

Donde a_i y b_i son los *i*-ésimos elementos en los conjuntos [A] y [B], w_a y w_b son posibles factores de ponderación relativos a los elementos del conjunto [A] y [B], respectivamente. A su vez, los indicadores compuestos pueden ser clasificados en función de su propósito y de la similitud entre las unidades de análisis:

  - o *Indicadores específicos*: permiten comparar aspectos similares entre unidades de distinto tamaño que pertenecen a un mismo sistema. Sin embargo, no son apropiados para determinar o comparar la distribución entre sistemas de diferente tamaño o naturaleza al tomar como referencia un solo elemento del conjunto seleccionado. Un prototipo de indicador específico es el factor de impacto.

  - o *Indicadores de equilibrio*: caracterizan las relaciones *input/output* en un sistema. Un ejemplo de este tipo de indicador es el *influence weight* empleado para el cálculo del nivel de investigación (*research level*).

  - o *Indicadores de distribución*: muestran aportaciones o contribuciones de un subsistema al total que es el estándar de referencia. Permiten caracterizar la contribución o peso que un rasgo o característica concreta representa. Aquí pueden citarse como ejemplos: la proporción de documentos según el tipo de

colaboración, el porcentaje de documentos altamente citados (*highly cited papers*) o el índice de actividad (*activity index*). Son muy utilizados para describir la actividad científica de organismos o países.

- o *Indicadores relativos*: posibilitan la comparación entre conjuntos (investigadores, equipos, instituciones) que presentan distintos rasgos bibliométricos como, por ejemplo, pertenecer a distintas áreas o especialidades. Tradicionalmente, como estándar de referencia absoluto se emplean medias o medianas. Un ejemplo de indicador relativo es el ratio de citas relativo (*relative citation rate)*.

- ✓ **Indicadores compuestos o *composite indicators***: se trata de indicadores que contienen dos o más indicadores parciales.

$$COI = \sum_{i=1}^{N} \frac{x_i}{\sum_{i=1}^{T} x_i} w_i$$

Donde COI es el valor del indicador compuesto para una institución, T es el número total de organismos evaluados, x_i es el valor *i*-ésimo del indicador parcial para esa institución, w_i es el factor de ponderación para el *i*-ésimo indicador parcial, y N es el número de indicadores parciales utilizados. Tratan de integrar en una sola medida, indicadores que describen diversos aspectos de la actividad científica. Un ejemplo de este tipo de indicador es el *general performance index* planteado por Vinkler (1998).

2.4.2. Indicadores según las dimensiones de la actividad científica

La siguiente clasificación aúna los distintos indicadores bibliométricos empleados en las publicaciones científicas que conforman esta tesis doctoral según la dimensión de la actividad científica que miden. No obstante, se incluyen algunas nociones e indicadores complementarios con objeto de ofrecer una visión más clara de su tipología y contexto de aplicación.

- ✓ **Indicadores de actividad científica:** se trata habitualmente de medidas basadas en recuentos de publicaciones con las que se pretende cuantificar los resultados o el rendimiento científico de distintos agregados. Su mayor utilidad surge para analizar la evolución temporal de la productividad o como elemento comparativo entre distintas unidades. Para esto último, como ya se ha señalado, es necesario establecer un estándar de referencia en el que ubicar los resultados.

- *Número de documentos/artículos*: mide el número de publicaciones científicas pertenecientes a un autor, centro de investigación, organismo, región, etc. Consisten en recuentos en bruto de documentos recogidos en las bases de datos *Web of* Science (*Science Citation Index Expanded*, *Social Sciences Citation Index* y *Arts & Humanities Citation Index*). Para analizar el desempeño científico y la relación de la producción con otros factores se emplea únicamente el grupo que constituyen los "ítems citables". Bajo esta agrupación se incluyen principalmente artículos originales, revisiones y *proceedings papers*[11] (en adelante referidos como artículos). Notas y cartas también pueden ser computadas, aunque sus tasas de citación son generalmente mucho menores.

- *Índice de Actividad o Activity Index (AI)*: propuesto por Frame (1977) como una medida relativa de la contribución de una institución frente al sistema al que pertenece. Para determinar la producción científica de una institución con respecto de todo un sistema, el *AI* puede definirse como:

$$AI = \frac{P_t}{P_T}$$

  Donde P_t denota la proporción de artículos que la institución dedica a un área (por ejemplo, la producción científica del CSIC), y P_T es la proporción que el campo o área representa en todo el sistema al que pertenece la unidad analizada (por ejemplo, la producción científica española). Un *AI > 1* denotaría una mayor especialización del CSIC en el tema que el promedio nacional y viceversa (González-Albo et al. 2012).

- ✓ **Indicadores de impacto observado**: se describen los indicadores derivados de las citas que reciben los artículos. En función de los métodos de normalización, las unidades de estudio y el aspecto concreto a analizar, pueden hallarse en la literatura una gran variedad de indicadores.

  - *Número de citas*: es el número absoluto de citas que recibe la producción de un investigador, centro, institución o país. Como en todos los indicadores relacionados con las citas es conveniente considerar la influencia de las autocitas.

[11] La tipología *proceedings papers* fue introducida en el año 2008 por la base de datos para identificar a los artículos que previamente habían sido presentado en un congreso. Dicha reclasificación se aplicó retrospectivamente al total de la base de datos. Actualmente, los artículos que han sido presentados en un congreso son designados simultáneamente como "artículo" y "proceedings paper" (González-Albo & Bordons, 2011). Para evitar duplicidades estos documentos se consideran como artículos.

- *Citas por documento/artículo*: representa el número medio de citas recibidas por los trabajos publicados por la unidad objeto de análisis. Ofrece una medida de la eficiencia de la citación. A diferencia de otros indicadores, como el indice-h, se ha observado que perjudica a los investigadores que fraccionan sus resultados para obtener un mayor número de publicaciones (Leimu & Koricheva, 2005).

- *Proporción de artículos no citados*: tomando como ejemplo un investigador, es el peso de los trabajos que no han recibido citas con respecto al total de artículos que ha publicado en el período objeto de estudio.

- *Ratio de Citas Relativo o Relative Citation Rate (RCR)*: permite mostrar si los artículos de la unidad de estudio atraen un mayor o menor número de citas que las del promedio internacional en su disciplina de publicación. También puede aplicarse a revistas. Al tratarse de un indicador estandarizado es posible comparar el impacto de la investigación publicada entre unidades de distinta naturaleza. De esta forma, las diferencias entre campos en el número absoluto de citas quedan prácticamente eliminadas (Schubert, Glänzel, & Braun, 1983).

$$RCR = \frac{MOCR}{MECR} = \frac{\sum_{i=1}^{n} c_i}{\sum_{i=1}^{n} e_i}$$

  Donde *MOCR* es el ratio medio de citas observado para la disciplina de estudio, *MECR* representa el ratio medio de citas esperado para la unidad de referencia, c_i denota el número de citas recibidas por el *i*-ésimo artículo de la unidad analizada, n es el número de artículos analizados, y e_i es el número de citas recibidas por el *i*-ésimo artículo que conforma el nivel de referencia superior. En las estudios donde se emplea, se ha considerado como unidad de referencia las citas medias recibidas por la disciplina a nivel mundial en el año de publicación del artículo. En el caso de revistas asignadas a más de una disciplina en la clasificación temática de *Web of Science*, se calcula un promedio. Un *RCR > 1* significa que la unidad de análisis ha recibido de media más citas que las del promedio del mundo en sus disciplina/s de publicación.

- *Mean Normalised Citation Score (MNCS)*: mide el número medio normalizado de citas que ha recibido, por ejemplo, un investigador.

$$MNCS = \frac{1}{n}\sum_{i=1}^{n}\frac{c_i}{e_i}$$

Donde n es el número de artículos que ha publicado el investigador en cuestión, c_i denota el número de citas recibidas por el artículo i y e_i representa el número medio de citas recibidas por los artículos publicados en la misma categoría, en el mismo año y que tienen la misma tipología documental que el artículo i (Waltman, van Eck, van Leeuwen, Visser, & van Raan, 2011). Las autocitas no se computan para el cálculo del indicador. Un valor del *MNCS = 1* significa que el investigador ha obtenido un número medio de citas igual al de la media mundial. Por lo tanto, si *MNCS > 1* quiere decir que el investigador ha obtenido en promedio más citas que la media mundial y viceversa.

Cabe señalar que el MNCS es el actual indicador de referencia o *crown indicator* empleado por el CWTS, sustituyendo al denominado CPP/FCSm (*Citation Per Publication/Field Citation Score mean*) que presentaba una metodología similar al RCR (Moed, De Bruin, & van Leeuwen, 1995). Ambos indicadores difieren en el mecanismo de normalización empleado. En el CCP/FCSm la normalización se realiza para el conjunto de artículos, es decir, primero se suma el número observado y esperado de citas y luego se divide uno por otro. En otras palabras, en el CPP/FCSm las publicaciones son entendidas como un todo, en vez de como una serie de trabajos independientes. Esto se traduce en que la distribución de las citas no se considera importante, por lo que los campos con mayor número de citas por artículo tienen mayor peso en el cálculo del indicador que aquellos con un menor número medio de citas por artículo. Este problema ha sido señalado y discutido en la literatura (ver por ejemplo, Opthof & Leydesdorff, 2010). En cambio, en el cálculo del MNCS todos los campos tienen el mismo peso sin importar el número medio de citas por artículo, es decir, primero se calcula el ratio de citas normalizado para cada artículo y después se obtiene un ratio de medias para el conjunto analizado. Por tanto, la normalización es llevada a cabo a nivel individual para las publicaciones que conforman el conjunto analizado. No obstante, aunque el MNCS es teóricamente más apropiado, parece que en la práctica las diferencias entre ambos indicadores no son excesivas (Waltman et al. 2011).

- *Ptop 10%*: este indicador cuantifica el número de artículos que, comparados con otros similares (misma disciplina, año de publicación y tipología documental), pertenecen al 10% de trabajos más citados. Se corresponde con el Percentil 90. A diferencia del MNCS o el RCR, que son indicadores independientes del tamaño muestral, el Ptop10% es un indicador en bruto para detectar la capacidad de, por ejemplo, un investigador para obtener trabajos altamente citados.

- ✓ **Indicadores relativos al prestigio de las revistas:** hacen referencia a la visibilidad e impacto de las revistas de publicación, tanto en valores absolutos como relativos.

Aquí se incluye el conocido factor de impacto[12], así como indicadores derivados donde se proponen distintos procedimientos de normalización para expresar más fidedignamente las diferencias entre revistas y áreas temáticas.

- *Factor de Impacto o Impact Factor*: propuesto por Garfield (1972), fue concebido inicialmente para su uso en bibliotecas y recuperación de información, para identificar las revistas más importantes dentro de una determinada disciplina. Sin embargo, ha evolucionado hasta convertirse en una herramienta muy utilizada en evaluación científica (Glänzel, Debackere, Thijs, & Schubert, 2006).

$$IF_y = \frac{c_y}{n_{y-1} + n_{y-2}}$$

Siendo IF_y el factor de impacto de una revista en el año y, c_y el número total de citas recibidas en el año y por todos los documentos publicados en la revista en los años $y-1$ e $y-2$, mientras que n_{y-1} y n_{y-2} son el número de artículos publicados por la revista dichos años (Vinkler, 2010).

Debido a su amplia difusión y utilización en el ámbito de la evaluación científica a nivel individual, se han observado situaciones de uso acrítico e inadecuado de este indicador que han llevado a diversos autores a poner manifiesto sus limitaciones: el factor de impacto de una revista no es representativo de todos los artículos publicados en ella, de hecho las distribuciones dentro de una misma revista suelen ser muy asimétricas (ver apartado 1.4.), habiéndose estimado que el 50% de los artículos concentran el 90% de las citas que obtiene la revista (Seglen, 1992); la diferente naturaleza y patrones de citación y envejecimiento de la literatura científica según las disciplinas, beneficia a aquellas con una vida media más corta debido a que la ventana de citación para el cálculo del indicador es de dos años[13]; la mayor parte de las autocitas de un artículo se concentran en los primeros años tras su publicación, por lo que tienen una alta influencia sobre el cálculo del factor de impacto (Glänzel et al., 2006). A pesar de estas y otras limitaciones, el factor de impacto resulta de interés pues permite identificar las revistas más importantes dentro de una disciplina. Estas revistas, además, cumplen con estrictos sistemas de selección de originales y procesos de revisión rigurosos, por lo que a pesar de que el factor de impacto no refleja el impacto real de un documento supone un

[12] Como indicador alternativo, la base de datos *Scopus* propuso el *Scimago Journal Rank* que presenta otras particularidades en su cálculo, como la utilización del algoritmo *PageRank* de *Google* (Pereira, Guerrero-Bote, & Moya-Anegón, 2010).

[13] Para reducir este problema *Thomson Reuters* calcula un factor de impacto con una ventana de citación de 5 años a partir del 2007.

indicador interesante sobre el prestigio y visibilidad de la investigación publicada (Bordons & Zulueta, 1999).

- *Primer Cuartil (Q_1)*: este indicador considera la proporción o porcentaje de artículos publicados por la unidad de estudio en cuestión en el 25% de revistas con mayor factor de impacto en cada disciplina[14]. En el caso de revistas asignadas a más de un disciplina se selecciona aquélla en la que ocupan una mejor posición.

- *Posición Normalizada de una Revista o Normalised Journal Position (NJP)*: propuesto por Bordons y Barrigón (1992), proporciona una medida que permite establecer la posición relativa de una revista en su disciplina científica, lo que posibilita comparar revistas de distintas disciplinas. La posición normalizada puede definirse como:

$$NJP = 1 - \frac{r_J}{J}$$

Donde r_J es el rango de la *i*-ésima revista y J es el número total de revistas de la disciplina. Los valores de la posición normalizada oscilan entre 0 y 1, de forma que valores próximos a la 1 indican una buena situación de la revista dentro de la disciplina.

- *Mean Normalised Journal Score (MNJS)*: es la media de citas normalizada de las revistas en las que publica una unidad calculada utilizando el mismo procedimiento que en la obtención del MNCS. Así, para la unidad objeto de estudio el MNJS es el promedio normalizado de citas recibidas por las revistas en las que ha publicado, en comparación con la media mundial para la misma disciplina, año y tipología documental.

- ✓ **Índice-h y derivados**: a raíz de la aparición del índice-h y del creciente interés por los estudios a nivel micro, ha emergido en la comunidad bibliométrica una copiosa bibliografía dedicada a discutir su base matemática, así como posibles normalizaciones y variantes (Alonso, Cabrerizo, Herrera-Viedma, & Herrera, 2009). Este indicador y sus variantes combinan dos dimensiones de la actividad científica, una cuantitativa (producción) y otra cualitativa (citas).

  - *Índice-h*: siguiendo la definición de Hirsch (2005), un científico tiene un indice-h si *h* de sus N_p artículos tienen al menos *h* citas cada uno, y el resto de artículos

[14] Desde una perspectiva estadística, resulta evidente que el 25% de revistas con mayor factor de impacto tienen por debajo al 75% de la distribución, es decir, se correspondería con el Q_3.

(N_p – h) tienen ≤ h citas cada uno. En la Figura 2.1 se presenta un ejemplo visual para su interpretación.

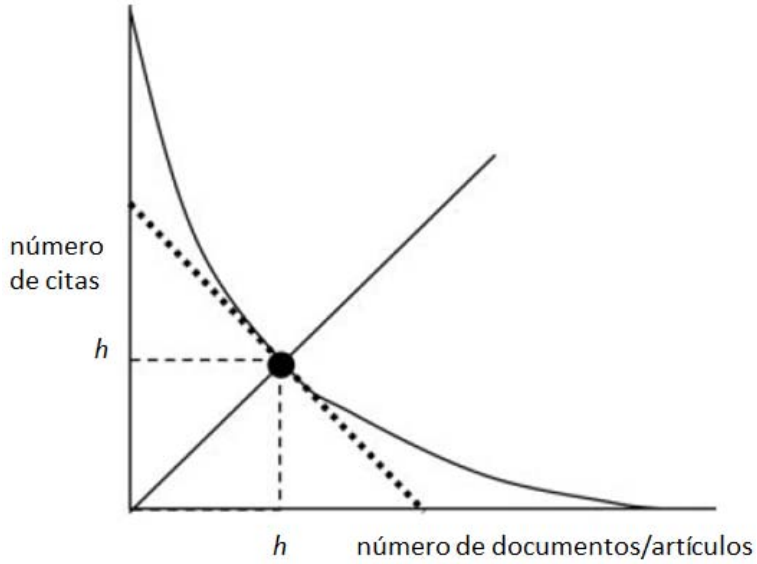


Figura 2.1. Interpretación gráfica del índice-h (fuente: Hirsch, 2005).

Como se muestra en la curva, que representa el número de artículos publicados por un investigador en orden decreciente según el número de citas recibidas, el índice-h es el punto donde el número de orden coincide con el número de citas recibidas (o no lo supera). El número total de citas es el área bajo la curva. Debido a su facilidad de cálculo, objetividad y a que no se ve afectado por distribuciones muy sesgadas (trabajos no citados o con tasas muy bajas de citas), se ha convertido en un indicador muy popular. Sin embargo, aunque los trabajos altamente citados (aquellos por encima de la intersección) son importantes para la determinación del índice-h, al trabajar con rangos de orden la magnitud de las citas no es tenida en cuenta (Egghe, 2006). Dado que el valor del índice-h para un determinado investigador no puede ser superior a su número de documentos, esto tiene como consecuencia que el indicador favorece a los investigadores muy productivos frente a aquellos más selectivos (Costas & Bordons, 2008).

- *Índice-g*: teniendo en cuenta la distribución altamente sesgadas de las citas, Egghe (2006) propuso el índice-g. Dado un conjunto de artículos ordenados en orden decreciente según el número de citas recibidas, el índice-g es el número (único) más alto de artículos que han recibido g^2 o más citas. Por tanto, en todos los casos $g \geq h$. Un ejemplo comparativo entre el valor obtenido para cada indicador se presenta en la Tabla 2.4.

Tabla 2.4. *Ranking* de los trabajos de Leo Egghe de acuerdo al número de citas recibidas (Egghe, 2006).

TC	*r*	$\sum TC$	r^2
47	1	47	1
42	2	89	4
37	3	126	9
36	4	162	16
21	5	183	25
18	6	201	36
17	7	218	49
16	8	234	64
16	9	250	81
16	10	266	100
15	11	281	121
13	12	294	144
13	13	307	169
13	14	320	196
13	15	333	225
12	16	345	256
12	17	357	289
12	18	369	324
12	19	381	361
11	20	392	400

TC representa el número citas de cada artículo con rango *r* y $\sum TC$ es el número de citadas acumuladas de los artículos de rango 1,2…,*r*. Las dos primeras columnas son necesarias para el cálculo del índice-h y las tres últimas para el índice-g. Las celdas sombreadas explican el valor obtenido para cada indicador. Como se puede observar para este autor $h = 13$, pues 13 de sus trabajos tienen al menos 13 citas, en consecuencia sus artículos con $r \geq 14$ han recibido no más de 13 citas. En cambio, $g = 19$ pues es el número mayor de artículos que han recibido al menos $19^2 = 361$ citas ($\sum TC = 381$), ya que para $r = 20$, $\sum TC < r^2$.

- ✓ **Indicadores de interdisciplinariedad:** la multi-asignación de revistas a categorías temáticas ha sido utilizada por algunos autores como aproximación al concepto de interdisciplinariedad (Morillo, Bordons, & Gómez, 2001). Otras aproximaciones que combinan la diversidad de categorías temáticas en las referencias de los documentos, índices de diversidad y coherencia, y análisis redes sociales han demostrado una mayor capacidad para captar la diversidad y heterogeneidad de los campos científicos (Rafols & Meyer, 2010); no obstante, presentan mayores dificultades para su obtención y tratamiento. En esta tesis doctoral la medición de la interdisciplinariedad se lleva a cabo a través del índice de Pratt.

- *Índice de Pratt*: es una medida que indica el nivel de concentración o dispersión de un conjunto de categorías temáticas a las que han sido asignadas un conjunto de ítems (Pratt, 1977). Este índice de concentración puede definirse de acuerdo a la siguiente fórmula:

$$PI = \frac{2\left(\frac{s+1}{2} - q\right)}{s-1}$$

con

$$q = \sum ia_i / t$$

Donde s representa el número de categorías temáticas de las publicaciones, a_i es el tamaño de la categoría de rango i y t es el número total de publicaciones del conjunto en cuestión. El rango de valores del índice oscila entre $PI = 0$, lo que se corresponde con una distribución uniforme de los documentos en las categorías temáticas, y $PI = 1$ donde todos los documentos pertenecerían exactamente a una categoría. En otras palabras, a menor valor del indicador mayor es el nivel de interdisciplinariedad. Tal y como señaló Carpenter (1979), el índice de Pratt supone simplemente una pequeña variante (normalización) con respecto al coeficiente de Gini, de mayor difusión y tradición, especialmente en el campo de la economía. De hecho, ambos indicadores difieren únicamente en el denominador, tomando Pratt valores de $s - 1$, mientras que en el de Gini el valor es igual a s.

- ✓ **Nivel de Investigación:** también es posible caracterizar la orientación básica o aplicada de la investigación a través de la clasificación de revistas propuesta por Narin, Pinski y Gee (1976). Originalmente fue ideada para describir la estructura de conocimiento en el ámbito biomédico, partiendo de la clara diferenciación existente entre el sector básico (investigación biomédica) y el sector aplicado (medicina clínica). Esta clasificación fue descrita por *CHI Research/Computer Horizons Inc.* que en la actualidad opera como ipIQ. La clasificación consta de 4 niveles que oscilan desde el nivel 1 (el más aplicado), hasta el 4 (el más básico). Para la asignación de revistas a un nivel de investigación, los autores emplearon revisión de expertos y tuvieron en cuenta los patrones de citación entre las revistas (*influence weight)*. La idea subyacente al estudio de los patrones de citación es que es posible apreciar una clara tendencia a citar revistas que pertenecen al mismo nivel que la que se estudia, pero también a las de niveles más básicos.

- ✓ **Indicadores de colaboración**: empleando la co-autoría como aproximación a la colaboración científica es posible cuantificar las interacciones que tienen lugar entre los investigadores, y por ende entre centros o instituciones, en la creación de

nuevo conocimiento. Como ya se ha señalado, la identificación de la afiliación de los autores se lleva a cabo a través del campo *Address* de la bases de datos. Por otro lado, también es posible emplear la información relativa a la posición de firma y el número de autores. De este modo, se emplean indicadores de distribución, específicos y relativos para caracterizar la colaboración científica de las unidades objeto de estudio.

- *Tasas de colaboración entre centros:* miden la contribución que cada tipo de colaboración representa respecto del total de la actividad científica de la unidad de estudio en cuestión. Se han definido tres tipos de colaboración: *colaboración internacional* (la que se produce instituciones de dos o más países), *colaboración nacional* (entre dos o más centros de un mismo país) y *sin colaboración* (pudiendo participar varios autores pero de una única institución, lo que no eximiría de colaboración a más bajo nivel, como la departamental). Suponiendo que *N* es número total de documentos y N_{intcol}, N_{natcol} y N_{nocol} son el número de documentos en colaboración internacional, nacional y con un único centro, respectivamente, la proporción que representa cada tipo de colaboración viene dada por:

$$P_{colint} = N_{intcol} \,/\, N$$

$$P_{colnat} = N_{natcol} \,/\, N$$

$$P_{nocol} = N_{nocol} \,/\, N$$

  En cuanto a aquellas publicaciones donde existe una colaboración mixta, es decir nacional e internacional, ésta se asigna a la colaboración internacional.

- *Índice de co-autoría*: este indicador mide el número medio de autores por documento en base a la producción total de la unidad analizada. Aunque se trata un promedio y carece de límite superior, se emplea este indicador pues permite caracterizar los patrones de colaboración entre autores dentro de una disciplina o especialidad. Proporciona una aproximación al tamaño medio de los grupos de investigación.

- *Posición de firma (primer autor, último autor)*: de forma similar a las tasas de colaboración, se calcula la proporción de trabajos publicados por un determinado autor en primera y última posición. Estas medidas permiten estimar la distribución y rol predominante de los autores en sus investigaciones.

- *Mean Geographical Collaborative Distance (MGCD)*: la distancia geográfica colaborativa se define como la mayor distancia geográfica entre las direcciones que recoge el campo *Address* de *Web of Science*. En el caso de que un artículo contenga únicamente una dirección, *MGCD = 0*. Para varias publicaciones se calcula un valor medio. Una descripción detallada del procedimiento de geo-codificación puede hallarse en Waltman, Tijssen y van Eck (2011).

- ✓ **Indicadores relacionales o de redes sociales**: una red social es un conjunto de individuos o grupos cada uno de los cuales tiene conexiones de algún tipo con todos o alguno de los integrantes de la red. El análisis de redes sociales asume que las relaciones o vínculos interpersonales son socialmente relevantes. Por ello, su principal objetivo es la detección e interpretación de estos vínculos sociales (De Nooy, Mrvar, & Batalegelj, 2011). El cálculo y aplicación de indicadores relacionales en los estudios de colaboración permite ahondar en la compleja estructura de relaciones e interacciones entre los agentes creadores de conocimiento.

  Una gran parte de los conceptos y características del análisis de redes sociales tienen su origen en la teoría de grafos. Tal y como señalan Wasserman y Faust (1994), la teoría de grafos ha resultado sumamente útil para el análisis de redes sociales pues ha proporcionado un vocabulario, que permite analizar distintas características de las estructuras sociales; operaciones matemáticas, para medir y describir estas propiedades; y teoremas, que pueden ser probados sobre los grafos permitiendo así contrastar hipótesis.

  En las redes, los individuos o grupos se conocen como actores o nodos (*nodes*) y las conexiones entre ellos como vínculos o lazos (*ties*). En las redes de colaboración, los actores son los autores de las publicaciones científicas y sus vínculos son medidos a través de la co-autoría. Una red de colaboración puede ser estudiada tanto a nivel macro (estructura de la red) como a nivel micro (patrones de relación entre los actores). Asimismo, en función del tipo de vínculo entre los nodos pueden distinguirse dos tipos de grafos: indirectos (relación recíproca entre los nodos, ambos nodos están involucrados igualmente en la relación) o directos (la relación no tiene por qué ser necesariamente recíproca, es decir, la relación es bidireccional). Debido a que los vínculos se establecen a través de la firma conjunta en las publicaciones, la relación entre los autores de una publicación es recíproca.

  Uno de los principales usos de la teoría de grafos en el análisis de redes sociales es la identificación de los actores más importantes, es decir, aquellos que ocupan posiciones de mayor relevancia o poder dentro de una red. Aquí, conviene distinguir entre dos conceptos, centralidad (*centrality)* y centralización (*centralization)*. Centralidad se refiere a la posición de los nodos dentro una red,

mientras que la centralización se emplea para caracterizar una red en su conjunto (De Nooy et al., 2011). Estos conceptos son multidimensionales, por lo que pueden ser medidos a través de distintos indicadores basados en el grado de una red (Freeman, 1979). Se describen a continuación las principales medidas de centralidad empleadas a nivel de autor:

- *Centralidad de Grado o Degree Centrality*: la idea de centralidad ha estado tradicionalmente vinculada a la de grado. El grado de un nodo p_i, es el número de actores p_j $(i \neq j)$ que son adyacentes a p_i y con los que, por tanto, mantiene un vínculo directo. Tomando como ejemplo un grafo de estrella (Figura 2.2a), se observa que el grado máximo que puede alcanzar cualquier actor en una red de cinco nodos es $C_D = 4$, valor que obtiene p_3. Esto implica que el actor p_3 ocupa una posición jerárquica y central en la red, donde el resto de actores son periféricos, es decir, están conectados con p_3 pero no tienen ningún vínculo entre ellos ($C_D = 1$). En cambio, en la red circular (Figura 2.2b) ningún actor ocupa una posición centralmente más ventajosa que el resto, pues el número vínculos es el mismo para los cinco nodos ($C_D = 2$).

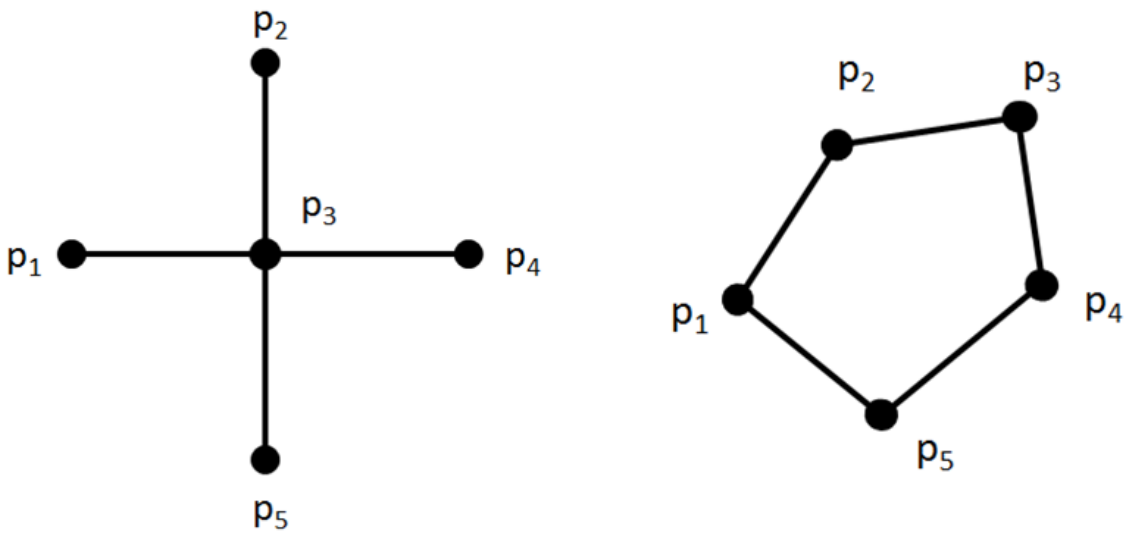


Figura 2.2. Redes de estrella (a) y circular (b) con cinco nodos.

Esta medida se limita a mostrar la centralidad de un actor, que en el caso de la colaboración científica es el número de colaboradores del nodo en cuestión (un investigador), y no es apta para comparar redes de distinto tamaño. Para eliminar el efecto del tamaño de la red y posibilitar las comparaciones redes de distinto tamaño, se obtiene una medida normalizada:

$$C'_D(p_k) = \frac{\sum_{i=1}^{n} a(p_i, p_k)}{n-1}$$

Donde C'_D es la centralidad de grado normalizada y n representa el número de nodos que componen la red. $a(p_i, p_k)$ = 1 si y solo si p_i y p_k han colaborado y es igual a 0 si no es así. De esta forma se obtiene una proporción que va desde 0 (nodo completamente aislado) a 1 (cuando el nodo está conectado con el resto de nodos de la red (Freeman, 1979).

- *Centralización de Grado o Degree Centralisation*: mientras la centralidad de grado es una medida a nivel individual, la centralización caracteriza a la red completa. Permite cuantificar la dispersión o variación de la centralidad entre los nodos de una red. Se calcula como la variación entre los grados de los actores dividido por el grado máximo posible en una red del mismo tamaño (Wasserman & Faust, 1994). Sus valores oscilan entre 0 que indica que todos los nodos tienen idéntica centralidad (red círculo) y 1 donde los actores colaboran únicamente con único nodo (red de estrella).

- *Centralidad de cercanía* o *Closeness Centrality*: proporciona una medida de lo "cerca" que está un actor de todos los demás actores de la red. En otras palabras, mide la capacidad de un nodo para acceder al resto de nodos de la red. Para ello, se tiene en cuenta el camino más corto entre dos nodos, es decir, el número de vínculos que los separan. Este camino más corto también se conoce como distancia geodésica. Así, la centralidad de un nodo es medida por su distancia geódesica a todos los demás nodos de la red. A mayor centralidad, menor es la distancia geodésica, es decir, menor número de vínculos necesita un investigador para relacionarse con el resto y viceversa. Por tanto, un nodo con una puntuación de cercanía muy baja ha de ser capaz influir directa o indirectamente sobre muchos otros (Borgatti, 2003; De Nooy et al., 2011).

$$C_C\ (p_k) = \sum_{i=1}^{n} d(p_i, p_k)^{-1}$$

Donde $d(p_i, p_k)$ es la distancia geodésica más corta entre p_i y p_k. Al igual que para la centralidad de grado, se puede obtener un valor normalizado dividiendo la cercanía del nodo entre el número nodos que componen la red menos uno.

- *Centralidad de Intermediación* o *Betweenness Centrality*: la intermediación es el número de distancias geodésicas, entre todos los pares de nodos, que pasan a través de un nodo dado. Por definición, en un grafo indirecto con p nodos, hay al menos $p\ (p-1)$ geodésicas.

$$C_B(p_k) = \sum_{i<}^{n} \sum_{j}^{n} \frac{g_{ij}(p_k)}{g_{ij}}, i \neq j \neq k$$

Donde g_{ij} es el número de geodésicas que unen al nodo p_i y al nodo p_j, mientras que $g_{ij}(p_k)$ denota las geodésicas que vinculan al nodo p_i y al nodo p_j y que pasan por el nodo p_k. Asimismo se puede obtener un valor normalizado:

$$C'_B(p_k) = \frac{2C_B(p_k)}{(n^2 - 3n + 2)}$$

Un nodo con alta intermediación une muchos pares de nodos a través del camino más corto, es decir, ese actor desempeña un papel de *gatekeeper* o *broker* y es probable que controle los flujos de información en la red. Por tanto, ocupa una posición estratégicamente ventajosa, pues si se eliminara dicho nodo muchos pares de nodos estarían más distantes o incluso aislados –como sucedería si se elimina el nodo p_3 en la red de estrella– (Freeman, 1979; Borgatti, 2003). El rango de valores oscila de 0 (un nodo se encuentra en todas las geodésicas de todos los pares de nodos) a 1 (un nodo no se encuentra en ninguna geodésica).

- <u>*Eigenvector Centrality*</u>: además del grado, cercanía e intermediación hay otra posible perspectiva de la centralidad. La asunción básica de la *eigenvector centrality* es que un nodo desempeña un papel más central en una red si los nodos con los que se vincula son a su vez centrales. Dicho de otro modo, es importante con quien colabora un investigador, pero también lo es con quienes colaboran sus colaboradores. Por tanto, el *eigenvector* como medida de centralidad supone asignar valores relativos a los nodos considerando que la centralidad de un nodo no depende únicamente del número de nodos adyacentes, sino también de la centralidad de esos nodos adyacentes (De Nooy et al., 2011). Así, un nodo altamente conectado con otros nodos que están a su vez bien conectados (con mayor grado), va a tener una alta centralidad de *eigenvector.* Se emplea el método de Kleinberg (1999) que es una medida próxima al poder de Bonacich (1972).

Además de las medidas de centralidad mencionadas, en la investigación realizada en este tesis se emplean algunos indicadores sobre la cohesión estructural de la redes, es decir, medidas que estiman la fortaleza de los vínculos entre los nodos. Cabe señalar que no hay un acuerdo en la literatura sobre que estructura de red resulta más ventajosa. Coleman (1988) apunta que redes densas con vínculos fuertes son vitales para fomentar la confianza mutua. Otros autores (ver Burt,

2004), defienden que la abundancia de vínculos débiles es ventajosa al fomentar la intermediación y el flujo de conocimiento entre actores heterogéneos. Una medida muy utilizada para analizar la cohesión, y en la que se basan algunos de los indicadores utilizados, es la densidad, que estima la proporción de vínculos presentes en una red respecto al máximo número de vínculos posibles.

- *Fuerza de los Vínculos* o *Strenght of Ties*: para evaluar la fuerza de los vínculos entre un nodo p_i y un nodo p_j se calcula una media de los pesos w_{ij} de sus co-autorías. Esto supone dividir la suma de los pesos de los vínculos del nodo (el número de co-autorías del investigador) entre el grado del nodo (número de colaboradores del autor).

- *Constraint*: permite evaluar si la red de colaboradores se concentra directamente o indirectamente en un solo nodo. Se puede calcular de la siguiente forma (Burt, 2004):

$$C_{ij} = (p_{ij} + \sum_q p_{iq}\, p_{qj})^2, \text{ para } q \neq i, j.$$

  Si dos nodos (j y q) son adyacentes al nodo i, la *constraint* (CO_{ij}) del nodo i por el nodo j se calcula como la suma de los cuadrados de la fuerza de los vínculos directos e indirectos del nodo i al nodo j. Se trata de una medida sobre la redundancia de contactos, es decir, mide como de abierta o cerrada es una red. Si un investigador tiene vínculos con otros que están a su vez altamente conectados a otros autores, el investigador en cuestión tiene muchos contactos redundantes y un alto valor para la *constraint.* Este tipo de autor puede estar malgastando sus recursos pues siempre crea y difunde el mismo tipo de información, en cambio, un autor sin colaboraciones redundantes puede que tenga acceso a una mayor diversidad de información, lo que se asocia a una investigación más novedosa (Abbasi, Altmann, & Hossain, 2011).

- *Coeficiente de Agrupamiento o Clustering Coefficient*: mide en qué proporción cada actor en una red está "integrado" en un cluster local. Para un nodo indica como de densas son las relaciones de su círculo de colaboradores. Puede definirse como la probabilidad de que dos colaboradores de un nodo sean adyacentes entre sí, es decir, de que hayan publicado un trabajo en colaboración (Abbasi et al., 2011). El coeficiente de agrupamiento, CCO, para un nodo i puede definirse como (Barabási et al., 2002):

$$CCO_i = 2n_i / k_i(k_i - 1)$$

Donde el nodo i tiene k_i vínculos con otros nodos de la red que forman un cliqué completo[15], hay $k_i(k_i - 1)$ vínculos entre los nodos y n_i es el número de vínculos que conectan a los k_i nodos seleccionados entre sí. Valores de agrupamiento cercano a la unidad muestran una elevada tasa de relaciones entre los colaboradores del nodo (alta probabilidad de que sus colaboradores publiquen juntos). Sin embargo, valores muy próximos a uno señalan que el nodo es el único vínculo entre sus colaboradores.

[15] Un cliqué consiste en un subconjunto de al menos tres nodos, todos adyacentes entre sí, de modo que todos los pares de nodos están directamente conectados a través de al menos una arista (Wasserman & Faust, 1994). Por tanto, un cliqué es una subred con máxima densidad.

3. Asunciones de partida y preguntas de investigación

En los capítulos precedentes, se ha discutido el papel de la bibliometría como herramienta para el análisis de la ciencia, se han explicado los orígenes y el rol primordial de la colaboración en la ciencia actual, su relación con otros factores, así como la naturaleza multidimensionalidad de la ciencia. Debido a las características de los datos bibliométricos, se ha puesto de manifiesto el interés de una aproximación multivariante y se ha presentado una revisión de las principales técnicas empleadas en el campo. Además, se han expuesto una serie de aspectos inherentes a todo estudio bibliométrico: la fuente de datos empleada, el tratamiento y normalización de los datos, una serie de consideraciones básicas a tener en cuenta, y sendas clasificaciones que detallan las características de los indicadores bibliométricos empleados.

Por tanto, del análisis de estos elementos se deducen una serie de conclusiones preliminares:

- ✓ En la actualidad, se considera que la colaboración es vital para el desarrollo de la investigación en la mayor parte de las áreas y campos científicos. Por un lado, se abordan problemas cada vez más complejos que requieren una aproximación interdisciplinar. Por otro lado, la disponibilidad de recursos es limitada y existe una fuerte dependencia de la financiación para poder llevar a cabo las investigaciones. No es de extrañar por tanto, que la investigación colaborativa sea fomentada por parte de las agencias financiadoras ya que permite compartir y optimizar el uso de los recursos, desarrollar aproximaciones multidisciplinares y se asocia a investigación de mayor calidad. La colaboración entre instituciones de distintos países constituye uno de los ejes centrales de los proyectos de la UE. En España, el Plan Nacional y los Planes propios de las distintas comunidades autónomas promueven la colaboración a distintos niveles (individuos, grupos, centros) y mediante distintos mecanismos (redes de investigación, consorcios, etc.).

- ✓ Necesidad de un análisis integrado de indicadores para abordar el estudio de la colaboración desde una perspectiva multidimensional. Se ha señalado en diversos estudios la influencia positiva de la colaboración sobre el impacto, la productividad o la interdisciplinariedad. Sin embargo, el grado de asociación puede variar según las áreas, disciplinas y/o países. En este sentido, las técnicas de análisis multivariante surgen como herramientas de sumo interés para estudiar las relaciones entre los diferentes indicadores bibliométricos, entre los indicadores y las unidades analizadas, así como las diferencias por países, áreas y disciplinas. Se ha observado un uso creciente en el campo bibliométrico de las técnicas de análisis multivariante, predominando la aplicación de MDS, Análisis Factorial con solución en Componentes Principales y el Análisis de Cluster.

- ✓ Importancia de los estudios a nivel micro, que generan mucho interés en el ámbito de la evaluación y la política científica, aunque presentan una mayor complejidad. Por un lado, los resultados son más sensibles a errores, como los derivados de la falta de normalización y, por otro lado, la dificultad para obtener grandes muestras hace que los resultados puedan presentar menor validez estadística. No obstante, resultan una unidad de estudio muy interesante al poder considerarse una gran variedad de indicadores y factores relacionados con el entorno de los investigadores. En este sentido, es importante incluir variables personales o asociadas al contexto tales como la edad, la categoría profesional o el sexo.

- ✓ La aproximación clásica de la bibliometría al estudio de la colaboración está basada en el análisis de la co-autoría. En la literatura se ha señalado que al equiparar co-autoría a colaboración se pueden pasar por alto algunas actividades colaborativas. En este sentido, tiene interés el estudio de otras fuentes de información sobre colaboración como es la sección de agradecimientos de los artículos científicos, que puede ser considerada una fuente de información sobre sub-autoría. La reciente inclusión de esta información en la base de datos *Web of Science* abre nuevas vías para los estudios sobre colaboración en la ciencia pues permite explorar otro tipo de interacciones más allá de los resultados que proporciona la co-autoría. Su explotación es un reto novedoso al tratarse de información no estructurada, es decir, texto en lenguaje natural.

A partir de estas observaciones preliminares, se plantean las preguntas de investigación objeto de este trabajo y a las que mediante los artículos publicados se trata de dar respuesta. Esta tesis se sustenta sobre la necesidad de una concepción multidimensional de la actividad científica y presenta nuevas aproximaciones metodológicas, a nivel meso y micro, desde la perspectiva de la estadística multivariante, especialmente a través de los métodos Biplot, y el análisis de redes sociales.

1) *¿Qué aportan los métodos Biplot, al estudio del papel de la colaboración en la actividad científica a nivel de centros e individuos?*

2) *¿Qué aportan las técnicas de análisis de redes sociales al estudio de la colaboración en la ciencia? ¿Cuál es la relación entre la posición de los autores en una red y su rendimiento científico?*

3) *¿Se puede detectar sub-autoría en las publicaciones científicas a partir del campo "agradecimientos"? ¿Existen diferencias por disciplina en la presencia del campo y en el tipo de colaboración predominante?*

4. Conclusiones generales e investigación futura

En este capítulo se exponen las conclusiones generales y se discute de forma global los resultados de las publicaciones que se presentan en la Parte 2. Por un lado, se pone de manifiesto el valor de las aportaciones metodológicas realizadas, por otro lado, se analiza el alcance de los resultados en el contexto de la política científica y la gestión de la investigación, y se trazan posibles líneas de investigación futuras.

Desde una perspectiva metodológica

Métodos Biplot y análisis multivariante: los métodos Biplot emergen como valiosas herramientas multivariantes para el análisis de matrices de datos bibliométricas, habiendo demostrado capacidad para inspeccionar las estructuras subyacentes de los datos y las relaciones entre las unidades de estudio y distintos conjuntos de indicadores.

A nivel de centros, el HJ-Biplot ha ofrecido una solución factorial óptima para describir la actividad de las áreas en lo que respecta a sus patrones de colaboración e impacto, e identificar aquellos centros con un comportamiento sobresaliente o singular que los diferencia del resto de su área. Si la técnica seleccionada para este fin hubiese sido un Análisis de Correspondencias, que únicamente permite trabajar con frecuencias, no se habría conseguido una solución adecuada al objetivo del estudio, debido a que esta técnica emplea la distancia ji-cuadrado. Esto significa que el peso asignado a cada fila/columna hubiese sido inversamente proporcional a su total marginal, es decir, se habría obtenido una representación simultánea de filas y columnas donde los centros con mayores valores de impacto y colaboración tendrían menor relevancia en el análisis y vendrían peor representados. Esto no sucede en el HJ-Biplot donde las masas son unitarias. En cambio, la aplicación de un Análisis de Correspondencias ha demostrado ser útil para identificar distintos patrones de agradecimientos por disciplina. Debido a que la base de datos *Web of Science* incluye únicamente esta información en el caso de mención expresa a la financiación, las palabras con mayores frecuencias estaban relacionadas con la financiación mientras que aquellas palabras que reflejaban algún tipo de deuda intelectual, apoyo o asistencia técnica presentaron frecuencias menores pero eran las relevantes para la búsqueda de información textual sobre colaboración científica (sub-autoría).

Por otro lado, la aplicación del Biplot Canónico, a una gran selección de indicadores que describen el complejo ecosistema a nivel individual, ha hecho posible determinar qué combinación de variables separa a los investigadores del CSIC agrupados por área y categoría profesional. A diferencia del HJ-Biplot, que busca las direcciones de máxima variabilidad, el Biplot Canónico construye una representación ponderada de la matriz

de medias que consigue las direcciones con máximo poder discriminante entre los grupos. Frente a otras técnicas de propósito similar como el MANOVA, el Biplot Canónico ofrece además la ventaja de proporcionar un gráfico factorial en dimensión reducida para la inspección visual de la matriz de datos. Si se hubiese empleado un Análisis Discriminante, también se obtendría un gráfico factorial en dimensión reducida, pero no se dispondría de información directa sobre los indicadores bibliométricos responsables de la separación entre las agrupaciones y sus correlaciones.

Estas consideraciones ponen de manifiesto la necesidad de tener un conocimiento pormenorizado de las características de los distintos métodos de análisis multivariante para identificar las técnicas adecuadas en cada caso y posibilitar la obtención de resultados pertinentes que respondan a los objetivos de estudio. Esta necesidad, aunque señalada hace casi tres décadas por Tijssen y de Leeuw (1988), resurge con especial importancia en el panorama actual en que el análisis multivariante está asentándose de forma definitiva entre las metodologías de análisis de la comunidad bibliométrica, tal y como ha puesto de manifiesto la revisión bibliográfica realizada. De hecho, si consideramos que los llamados 'mapas de la ciencia', que en ocasiones incluyen metodologías multivariantes, han sido propuestos como instrumentos de apoyo en política científica y gestión de la investigación (Noyons, 1999; 2001), la necesidad de incorporar estos conocimientos cobra un significado aún mayor.

Análisis de redes sociales: aunque empleado sobre todo en bibliometría para visualizar la estructura de los campos científicos ("mapeo"), se presenta en esta tesis una aproximación que va más allá de la representación gráfica de las redes y se basa en el uso de distintas medidas de centralidad y cohesión para describir los campos. Siguiendo los trabajos de Abbasi y colaboradores (Abbasi et al. 2011; Abbasi, Chung, & Hossain, 2012), esta metodología nos ha proporcionado medidas para describir de forma global la estructura de una disciplina en función de sus redes de co-autoría (macro); indicadores para caracterizar la posición y rol de cada uno de los autores que integran una red (micro); y la posibilidad de incluir dichas medidas en un modelo de regresión para evaluar la influencia de las mismas sobre el desempeño científico. Esta aproximación asume que la posición de un autor dentro de una red determina en cierta medida sus oportunidades y limitaciones, y que, por tanto, ejerce una influencia importante sobre los resultados que va a ser capaz de producir (Borgatti, Mehra, Brass, & Labianca, 2009). Por consiguiente, ahondar en las estructuras de producción de conocimiento y tratar de asociar ciertos roles o posiciones a un mejor desempeño científico (uso combinado de análisis de redes sociales e indicadores bibliométricos), proporciona información que va más allá de la que se extrae del análisis de los vínculos directos que mantienen los investigadores (co-autoría) y permite discernir la importancia relativa de éstos dentro del círculo social en el que desarrollan su actividad científica.

Agradecimientos como fuente de información sobre colaboración científica: se confirma que la sección de los agradecimientos contienen información social, cognitiva e instrumental sobre el proceso científico. En concreto, se valida la hipótesis sugerida por Patel (1973) y Heffner (1981) que afirmaba que los agradecimientos podían ser considerados como una fuente de información sobre sub-autoría, pero que hasta la reciente inclusión de esta información en *Web of Science* no se había podido comprobar a una escala razonable. A pesar de que los resultados pueden ser considerados como un estudio piloto, se ha identificado el diferente rol desempeñado por los agradecimientos en función de la disciplina. Por tanto, la explotación e inclusión sistemática de esta información junto a la aproximación bibliométrica clásica basada en la co-autoría, permitiría subsanar en parte alguna de las limitaciones señaladas en la literatura al equiparar co-autoría a colaboración científica (Melin & Persson, 1996; Laudel, 2002), y ofrecer una imagen más nítida y fidedigna de las interacciones que tienen durante el proceso colaborativo.

Desde la perspectiva de la política científica

Profundizar en el conocimiento del proceso científico e identificar qué factores se asocian a un mejor desempeño es importante para apoyar el desarrollo de políticas científicas adecuadas, tanto a nivel institucional como regional o nacional. El rasgo distintivo de la investigación analizada en esta tesis ha sido la existencia de diferentes patrones de colaboración según las áreas y la asociación de la colaboración, especialmente la internacional, con un mejor desempeño científico. A nivel de área, se ha evidenciado en el análisis de los centros del CSIC una correlación positiva entre el impacto de la investigación y la colaboración internacional, aunque la incidencia de esta última es variable según las áreas. Por otro lado, se ha observado cierto grado de heterogeneidad dentro las áreas y la posibilidad de identificar centros atípicos que se alejan del patrón dominante de su área, lo que en algunos casos se explica por la naturaleza de la investigación realizada (carácter básico/aplicado, orientación local/internacional, etc.). En definitiva, se pone de manifiesto la importancia de favorecer la colaboración internacional, pero también de tener en cuenta las características de cada área con objeto de establecer pautas de actividad, estudio y evaluación coherentes con el tipo de investigación que se desarrolla en cada caso.

Los estudios micro pueden ser especialmente relevantes a nivel de centros, pues permiten profundizar en el desempeño científico de los investigadores y en sus prácticas de colaboración, así como explorar la influencia de factores personales y organizativos. El estudio micro de dos áreas CSIC ha puesto de manifiesto que los investigadores presentan distintos patrones de actividad según su categoría profesional y área de adscripción. Se ha observado que durante la etapa post-doctoral los investigadores asumen el trabajo experimental y firman sobre todo en primer

lugar; mientras que al ascender en la escala científica tienden a tener una red de colaboradores más diversa, mayores niveles de producción y un mayor papel de liderazgo y supervisión, que se refleja en su tendencia a firmar en último lugar y que se ha relacionado con su función como fuerza motriz de los grupos de investigación (Liu & Fang, 2014).

En cambio, aunque los resultados en términos de citas y calidad de las revistas donde publican los investigadores son similares para las distintas categorías, se puede afirmar que los investigadores post-doc difunden su trabajo a través de revistas de mayor prestigio que científicos en escalas superiores. Este resultado puede reflejar los efectos que las políticas científicas están teniendo sobre los hábitos de publicación de jóvenes investigadores, conscientes de la importancia de publicar en revistas internacionales de prestigio como un factor clave para la promoción y avance de sus carreras científicas. No obstante, aunque la publicación en revistas de alto prestigio es sin duda positiva para aumentar la visibilidad de la investigación, la competitividad impuesta por un sistema de promoción basado en la publicación puede tener consecuencias negativas sobre los hábitos colaborativos de los científicos. En este sentido, aflora como una línea de investigación sugerente la tensión existente entre colaboración y competitividad, especialmente en las primeras etapas de las carreras académicas (van den Besselaar, Hemlin, & van der Weijden, 2012). Las instituciones financiadoras fomentan la colaboración a distintos niveles por considerar que es positiva para la investigación; pero, al mismo tiempo, recompensan la autoría única o en posición de máxima responsabilidad, lo que puede ser difícil de alcanzar en áreas muy colaborativas. Los datos del CSIC muestran la tendencia de los investigadores post-doctorales a firmar como primer autor en sus investigaciones, lo que puede incrementar sus oportunidades de promoción en el futuro. Este enfoque puede conducir a los científicos jóvenes a evitar co-autorías elevadas y disminuir sus relaciones con otros investigadores con objeto de asegurar una posición dominante en las publicaciones. De hecho, estos efectos colaterales han sido descritos en el ámbito de ciencias de la salud, observándose una preferencia de los científicos post-doctorales hacia la co-autoría individual como forma de evitar conflictos y asegurarse una posición relevante en la publicación (Müller, 2012).

El análisis de redes sociales a nivel individual ha revelado que la estructura y composición de las redes varía por disciplina, y que con independencia de ésta determinadas posiciones en las redes se asocian a un mejor desempeño científico. Así, aquellos investigadores con una red de colaboradores más amplia y/o que trabajan de forma estable y duradera con determinados científicos muestran mayores valores del índice-g; mientras que no se han observado beneficios asociados a otras posiciones como las de intermediación. Por consiguiente, compartir conocimientos y establecer relaciones duraderas construidas en base a la confianza mutua son aspectos catalizadores de una investigación de mayor calidad. Este dato sería coherente con la

menor necesidad de imponer mecanismos de coordinación y comunicación entre colaboradores de distintos países descrito para los grupos o redes de investigación internacionales consolidadas (Cummings & Kiesler, 2005). En este sentido, fomentar las redes de colaboración estables y duraderas (especialmente en áreas muy dependientes de la financiación y de grandes infraestructuras), que aúnan los beneficios encontrados a través de las distintas metodologías empleadas y minimizan algunos problemas asociados a la misma, permitiría impulsar la investigación hacia mayores cotas de excelencia.

Finalmente, se pone de manifiesto la utilidad de la sección de agradecimientos de las publicaciones en el campo de la política científica. Por un lado, porque incluye la mención de la fuente de financiación recibida para desarrollar la investigación, lo que sin duda interesa a las agencias financiadoras para realizar un seguimiento de sus ayudas. Por otro lado, recoge información sobre colaboradores no mencionados en la autoría. El análisis de los patrones de agradecimientos, realizado en esta tesis para cuatro disciplinas, podría extenderse a otras áreas y resultar en un "mapa de los agradecimientos" según áreas temáticas, que permitiría agrupar éstas según sus necesidades de colaboración. En cualquier caso, sería necesaria una mayor sistematización y normalización de la información incluida en el campo de agradecimientos para facilitar su tratamiento y análisis.

Referencias

Abbasi, A., Altmann, J., & Hossain, L. (2011). Identifying the effects of co-authorsip networks on the performance of scholars: a correlation and regression analysis of performance measures and social network analysis measures. *Journal of Informetrics*, *5*, 594-607.

Abbasi, A., Chung, K.S.K., & Hossain, L. (2012). Egocentric analysis of co-authorship network structure, position and performance. *Information Processing and Management*, *48*(4), 671–679.

Abd el Kader, M., Ojasoo, T., Miquel, J.F., Okubo, Y., & Doré, J.C. (1998). Hierarchical author networks: an analysis of European Molecular Biology Laboratory (EMBL) publications. *Scientometrics*, *42*(3), 405-421.

Abramo, G., D'Ángelo, C.A., & Di Costa, F. (2009). Research collaboration and productivity is there correlation? *Higher Education, 57*, 155-171.

Abramo, G., D'Angelo, C.A., & Di Costa, F. (2011). Research productivity: Are higher academic ranks more productive than lower ones? *Scientometrics*, *88,* 915–928. doi: 10.1007/s11192-011-0426-6

Abramo, G., D'Angelo, C.A., & Solazzi, M. (2011). Are researchers that collaborate more at the international level top performers? An investigation on the Italian university system. *Journal of Informetrics*, *5*, 204-211. doi: 10.1016/j.joi.2010.11.002

Albarrán, P., Crespo, J.A., Ortuño, I., & Ruiz-Castillo, J. (2011). The skewness of science in 219 sub-fields and a number of aggregates. *Scientometrics*, *88*(2), 385-397.

Alonso, S., Cabrerizo, F. J., Herrera-Viedma, E., & Herrera, F. (2009). H-Index: a review focused in its variants, computation and standardization for different scientific fields. *Journal of Informetrics*, *3*(4), 273-289.

Archambault, E., Vignola-Gagne, E., Coté, G., Lariviere, V., & Gingras, Y. (2006). Benchmarking scientific output in the social sciences and humanities: the limits of existing databases. *Scientometrics*, *68*(3), 329-342.

Barabási, A.L., Jeong, H., Ravasz, E., Néda, Z., Schuberts, A., & Vicsek, T. (2002). Evolution of the social network of scientific collaborations. *Physica A*, *311*, 590-614.

Batagelj, V., & Mrvar, A. (2013). Pajek. V 3.14. http://vlado.fmf.uni-lj.si/pub/networks/pajek/

Beaver, D. B. (1986). Collaboration and teamwork in physics. *Czechoslovak Journal of Physics*, *36*(1), 14-18.

Beaver, D. deB. (2001). Reflections on scientific collaboration (and its study): Past, present, and future. *Scientometrics*, *52*(3), 365–377.

Beaver, D. deB., & Rosen, R. (1978). Studies in scientific collaboration: Part I - the professional origins of scientific co-authorship. *Scientometrics*, *1*, 65-84.

Beaver, D. deB., & Rosen, R. (1979a). Studies in scientific collaboration: Part II - Scientific co-authorship, research productivity and visibility in the French scientific elite, 1799–1830. *Scientometrics*, *1*(2), 133-149.

Beaver, D. deB., & Rosen, R. (1979b). Studies in scientific collaboration: Part III - Professionalization and the natural history of modern scientific co-authorship. *Scientometrics*, *1*(2), 231–245.

Benzécri, J.P. (1973). *L'analyse de Données. Vol. 2. L'analyse des correspondances*. Paris: Dunod.

Birnholtz, J. (2006) What does it mean to be an author? The intersection of credit, contribution and collaboration in science. *Journal of the American Society for Information Science and Technology*, *57*(13), 1758–1770.

Bonacich, P. (1972). Factoring and weighting approaches to status scores and clique identification. *Journal of Mathematical Sociology*, *2*(1), 113-120.

Bordons, M., Aparicio, J., & Costas, R. (2013). Heterogeneity of collaboration and its relationship with research impact in a biomedical field. *Scientometrics*, *96*, 443-466. doi: 10.1007/s11192-012-0890-7

Bordons, M., & Barrigón, S. (1992). Bibliometric analysis of publications of Spanish pharmacologists in the SCI (1984-1989) Part I. *Scientometrics*, *25*(3), 425-446.

Bordons, M., & Gómez, I. (2000). Collaboration Networks in Science. In B. Cronin & H. B. Atkins (Eds.), *The web of knowledge. A festschrift in honor of Eugene Garfield* (197-213). Medford, NJ: ASIS Monograph.

Bordons, M., Morillo, F., Gómez, I., Moreno, L., Aparicio, J., & González-Albo, B. (2014). *La actividad científica del CSIC a través de indicadores bibliométricos (Web of Science 2009-2013)*. Madrid: IFS-UTAI-CCHS-CSIC. http://hdl.handle.net/10261/109251

Bordons, M., & Zulueta, M.A. (1999). Evaluación de la actividad científica a través de indicadores bibliométricos. *Revista Española de Cardiología*, *52*, 790-800.

Borgatti, S.P. (2003). The key player problem. In R., Breiger, K. Carley and P. Pattison (Eds.) *Dynamic Social Network Modelling and Analysis: Workshop Summary and Papers*, pp. 241-255. Washington, D.C.: National Academy Press.

Borgatti, S.P., Mehra, A., Brass, D.J., & Labianca, G. (2009). Network analysis in the social sciences. *Science*, *323*, 892–895.

Börner, K., Chen, C., & Boyack, K.W. (2003). Visualizing knowledge domains. In: Blaise Cronin (Eds.). *Annual Review of Information Science and Technology*, *37*, 179-255.

Bornmann, L., & Leydesdorff, L. (2014). Scientometrics in a changing research landscape. *Embo reports*, *15*(12), 1228-1232.

Bozeman, B., Fay, D., & Slade, C.P. (2013). Research collaboration in universities and academic entrepreneurship: the-state-of-the-art. *J Technol Transf*, *38*, 1-67. doi: 10.1007/s10961-012-9281-8

Bozeman, B., & Gaughan, M. (2011). How do men and women differ in research collaborations? An analysis of the collaborative motives and strategies of academic researchers. *Research Policy*, *40*, 1393-1402. doi: 10.1016/j.respol.2011.07.002

Bradford, S.C. (1948). *Documentation*. London: Crosby Lockwood.

Burt, R.S. (2004). Structural holes and good ideas. *American Journal of Sociology, 110*(2), 349-399.

Carayol, N., & Matt, M. (2004). Does research organization influence academic production? Laboratory level evidence from a large European university. *Research Policy*, *33*, 1081-1102.

Carpenter, M.P. (1979). Similarity of Pratt's measure of class concentration to the Gini index. *Journal of the American Society for Information Science*, *30*(2), 108-110.

Chen, C. (2006). CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature. *Journal of the American Society for Information Science and Technology*, *57*(3), 359–377.

Claxton, L.D. (2005). Scientific authorship part 2. History, recurring issues, practices, and guidelines*. Mutation Research*, *589*, 31–45.

Coleman, J.S. (1988). Social capital in the creation of human capital. American *Journal of Sociology*, *94*(Suppl.), 95-120.

Costas, R. (2008). *Análisis bibliométrico de la actividad científica de los investigadores del CSIC en tres áreas: biología y biomedicina, ciencia de materiales y recursos naturales: una aproximación metodológica a nivel micro (Web of Science, 1994-2004)* (Tesis doctoral no publicada). Universidad Carlos III, Departamento de Biblioteconomía y Documentación, Madrid, España.

Costas, R., & Bordons, M. (2008). Is g-index better than h-index? An exploratory study at the individual level. *Scientometrics*, *77*(2), 267-288.

Costas, R., van Leeuwen, T.N., & Bordons, M. (2010). A bibliometric classificatory approach for the study and assessment of research performance at the individual level : the effects of age on productivity and impact. *Journal of the American Society for Information Science and Technology*, *61*(8), 1564–1581. doi: 10.1002/asi.21348

Crane, D. (1972). *Invisible colleges: diffusion of knowledge in scientific communities*. Chicago: Chicago University Press.

Cronin, B. (1995). *The Scholar's Courtesy: The Role of Acknowledgments in the Primary Communication Process*. Los Angeles: Taylor Graham.

Cronin, B. (2001). Hyperauthorship: A postmodern perversion or evidence of a structural shift in scholarly communication practices? *Journal of the American Society for Information Science and Technology*, *52*(7), 558–569.

Cuadras, C.M. (2014). *Nuevos métodos de análisis multivariante*. Manacor (Mallorca): CMC Editions.

Cummings, J.N., & Kiesler, S. (2005). Collaborative research across disciplinary and organizational boundaries. *Social Studies of Science*, *35*, 703-722.

De Lange C., & Glänzel, W. (1997) Modelling and measuring multilateral coauthorship in international scientific collaboration. Part I. Development of a new model using a series expansion approach. *Scientometrics*, *40*(3), 593-604.

De Leeuw, J., & van der Heijden, P.G.M. (1985). *Quasi-Correspondence analysis*. (Research Report RR-85-19). Department of Data Theory, University of Leiden.

De Nooy, W., Mrvar, A., & Batagelj, V. (2011). *Exploratory social network analysis with Pajek: revised and expanded second edition*. Cambridge: Cambridge University Press.

Ding, Y. (2011). Scientific collaboration and endorsement: network analysis of coauthorship and citation networks. *Journal of Informetrics, 5*(1), 187-203. doi: 10.1016/j.joi.2010.10.008

Doré, J.C., Dutheuil, C., & Miquel, J.F. (2000). Multidimensional analysis of trends in patent activity. *Scientometrics*, *47*(3), 475-492.

Duque, R.B., Ynalvez, M., Sooryamoorthy, R., Mbatia, P., Dzogbo, D.B.S., & Shrum. W. (2005). Collaboration paradox: scientific productivity, the internet, and problems of research in developing. *Social Studies of Science*, *35*(5), 755-785

Egghe, L. (1988). Methodological aspects of bibliometrics. *Library science with a slant to documentation and information studies*, *25*, 179-191.

Egghe, L. (2005). *Power laws in the information production process: Lotkaian informetrics*. Kidlington: Elsevier Academic Press.

Egghe, L. (2006). Theory and practise of the g-index. *Scientometrics*, *69*(1), 131-152.

Egghe, L., & Rousseau, R. (1990). *Introduction to Informetrics: Quantitative methods in library, documentation and information science*. Amsterdam: Elsevier.

Fernández, M.T., Cabrero, A., Zulueta, M.A., & Gómez, I. (1993). Constructing a relational database for bibliometric analysis. *Research Evaluation*, *3*(1), 55-62.

Frame, J.D. (1977), Mainstream research in Latin America and the Caribbean. *Interciencia*, *2*, 143-148.

Franceschet, M., & Costantini, A. (2010). The effect of scholar collaboration on impact and quality of academic papers. *Journal of Informetrics*, *4*, 540-553. doi: http://dx.doi.org/10.1016/j.joi.2010.06.003

Freeman, L.C. (1979). Centrality in social networks: conceptual clarifications. *Social Networks*, *1*, 215-239.

Gazni, A., & Didegah, F. (2011). Investigating different types of research collaboration and citation impact: a case study of Harvard University's publications. *Scientometrics*, *87*, 251-265. doi: 10.1007/s11192-011-0343-8

Ganzi, A., Sugimoto, C.R., & Didegah, F. (2012). Mapping world scientific collaboration: authors, institutions, and countries. *Journal of the American Society for Information Science & Technology*, *63*(2), 323-335. doi: 10.1002/asi.21688

Garfield, E. (1972). Citation analysis as a tool in journal evaluation. *Science*, *178*, 471-479.

Giles, C.L., & Council, I.G. (2004). Who gets acknowledged: Measuring scientific contributions through automatic acknowledgment indexing. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(51), 17599–17604.

Glänzel, W. (2000). Science in Scandinavia: a bibliometric approach. *Scientometrics*, *48*(2), 121-150.

Glanzel, W. (2001). National characteristics in international scientific co-authorship relations. *Scientometrics*, *51*(1), 69-115.

Glänzel, W., Debackere, K., Thijs, B., & Schubert, A. (2006). A concise review on the role of author self-citations in information science, bibliometrics and science policy. *Scientometrics*, *67*(2), 263-277.

Glänzel, W., & Schubert, A. (2004). Analysing scientific networks through co-authorship. In: H.F. Moed, W. Glänzel, and U. Schmoch (Eds.) pp. 257-276, *Handbook of Quantitative Science and Technology Research*. Dordrecht: Kluwer Academic Publishers, 2004.

Glänzel, W., & Wouters, P. (2013). The dos and don'ts in individual level bibliometrics. *14th International Society of Scientometrics and Informetrics Conference (ISSI 2013)* (34 pp.). Viena: Universidad de Viena. <http://www.slideshare.net/paulwouters1/issi2013-wg-pw>

Gómez, I., Bordons, M., Morillo, F., Moreno, L., Aparicio, J., Candelario, A., González-Albo, B., & Herrero, M. (2009). *Indicadores de Producción Científica y tecnológica de la Comunidad de Madrid (2004-2008)*. Madrid: IEDCYT, CSIC.

González-Albo, B., & Bordons, M. (2011). Articles vs. proceedings papers: Do they differ in research relevance and impact? A case study in the Library and Information Science field. *Journal of Informetrics*, *5*, 369-381. doi:10.1016/j.joi.2011.01.011

González-Albo, B., Moreno, L., Morillo, F., & Bordons, M. (2012). Indicadores bibliométricos para el análisis de la actividad de una institución multidisciplinar: el CSIC. *Revista Española de Documentación Científica*, *35*(1), 9-37.

González-Alcaide, G., & Gómez-Ferri, J. (2014). La colaboración científica: principales líneas de investigación y retos de futuro. *Revista Española de Documentación Científica*, *37*(4), e062. doi: http://dx.doi.org/10.3989/redc.2014.4.1186

González-Pereira, B., Guerrero-Bote, V. P., & Moya-Anegón, F. (2010). A new approach to the metric of journals' scientific prestige: The SJR indicator. *Journal of Informetrics*, *4*(3), 379-391.

Guerrero-Bote, V.P., Olmeda-Gómez, C., & Moya-Anegón, F. (2013). Quantifying the benefits of international scientific collaboration. *Journal of the American Society for Information Science and Technology*, *64*(2), 392-404. doi: 10.1002/asi.22754

Gupta, B.M., Kumar, S., & Rousseau, R. (1998). Applicability of selected probability distributions to the number of authors per article in theoretical population genetics. *Scientometrics*, *42*(3), 325-334.

Haustein, S., & Larivière, V. (2015). The use of bibliometrics for assessing research: possibilities, limitations and adverse effects. *Incentives and Performance*, *45*, 121-139.

Heffner, A.G. (1981). Funded research, multiple authorship, and subauthorship collaboration in four disciplines. *Scientometrics*, *3*(1), 5–12.

Hirsch, J.E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(46), 16569–16572.

Hotelling, H. (1993). Analysis of a complex of statistical variables into principal components. Journal of Educational Psychology, *24*(6), 417-441.

Ibáñez, A., Larrañaga, P., & Bielza, C. (2013). Cluster methods for assessing research performance: exploring Spanish computer science. *Scientometrics*, *97*(3), 571-600.

ICMJE (2015). Uniform requirements for manuscripts submitted to biomedical journals: Ethical considerations in the conduct and reporting of research: Authorship and contributorship. http://www.icmje.org/ [Acesso: 16 de febrero, 2015].

Jarneving, B. (2008). A variation of the calculation of the first author cocitation strength in author cocitation analysis. *Scientometrics*, *77*(3), 485-504.

Jarneving, B. (2010). Regional research and foreign collaboration. *Scientometrics*, *83,* 295-320.

Jha, Y., & Welch, E. (2010). Relational mechanisms governing multifaceted collaborative behavior of academic scientists in six fields of science and engineering. *Research Policy*, *39*(9), 1174–1184.

Katz, J.S. & Hicks, D. (1997). How much is a collaboration worth? A calibrated bibliometric model. *Scientometrics, 40*(3), 541-554.

Katz, J.S., & Martin, B.R. (1997). What is research collaboration? *Research Policy*, *26*(1), 1-18.

Kleinberg, J.M. (1999). Authoritative Sources in a Hyperlinked Environment. *Journal of the ACM*, *46*(5), 604-632.

Kostoff, R.N. (1995). Federal research impact assessment - axioms, approaches, aplications. *Scientometrics*, *34*, 136-145.

Laudel, G. (2002). Collaboration and reward. What do we measure by co-authorships? *Research Evaluation*, *11*(1), 3–15.

Lee, S., & Bozeman, B. (2005). The impact of research collaboration on scientific productivity. *Social Studies of Science*, *35*, 673–702.

Leimu, R., & Koricheva, J. (2005). What determines the citation frequency of ecological papers? *Trends in Ecology and Evolution*, *20*(1), 438-443. doi: 10.1016/j.tree.2004.10.010

Lewison, G. (1999). Definition and calibration of biomedical fields. *Scientometrics*, *46*(3), 529-537.

Lewison, G., & Markusova, V. (2010). The evaluation of Russian cancer research. *Research Evaluation*, *19*(2), 129-144.

Leydesdorff, L., Kushnir, D., & Rafols, I. (2014). Interactive overlay maps for US patent (USPTO) data based on International Patent Classification (IPC). *Scientometrics*, *98*(3), 1583-1599.

Liu, X.Z., & Fang, H. (2014). Scientific group leaders' authorship preferences: an empirical investigation. *Scientometrics*, *98*, 909-925. doi: 10.1007/s11192-013-1083-8

Lotka, A.J. (1926). The frequency distribution of scientific productivity. *Journal of Washington Academy of Sciences*, *16*(12), 317-323.

Luukkonen, T., Persson, O., & Siversten, G. (1992). Understanding patterns of international scientific collaboration. *Science, Technology & Human Values*, *17*(1), 101-126.

Mabe, M. (2003). The growth and number of journals. *Serials*, *16*(2), 191-197.

Martin, B.R. (1996). The use of multiple indicators in the assessment of basic research. *Scientometrics 36*(3), 343-362.

Martin, B.R., & Irvine, J. (1983). Assessing basic research: some partial indicators for scientific progress in radio astronomy. *Research Policy*, *12*, 61-90.

Maltrás-Barba, B. (1996). *Los indicadores bibliométricos en el estudio de la ciencia. Fundamentos conceptuales y aplicación en política científica* (Tesis doctoral no publicada). Universidad de Salamanca, Facultad de Filosofía, Salamanca, España.

McCain, K.W. (1983). The author cocitation structure of macroeconomics. *Scientometrics*, *5*(5), 277-289.

Melin, G. (2000). Pragmatism and self-organization: research collaboration on the individual level. *Research Policy*, *29*, 31-40.

Melin, G., & Persson, O. (1996). Studying research collaboration using co-authorships. *Scientometrics*, *36*(3), 363–377.

Merton, R.K. (1968). The Matthew effect in science. *Science*, *159*, 56-63.

Miquel, J.F., Ojasoo, T., Okubo, Y., Paul, A., & Doré, J.C. (1995). World science in 18 disciplinary areas: Comparative evaluation of the publication patterns of 48 countries over the period 1981–1992. *Scientometrics*, *33*(2), 149-167.

Moed, H.F. (2000). Bibliometric indicators reflect publication and management strategies. *Scientometrics*, *47*(2), 323-346.

Moed, H.F. (2005). *Citation analysis in research evaluation.* The Netherlands: Springer.

Moed, H.F., De Bruin, R.E., & Van Leeuwen, T.N. (1995). New bibliometric tools for the assessment of national research performance: Database description, overview of indicators and first applications. *Scientometrics*, *33*(3), 381–422.

Moore, G., & Shiskin, J. (1967). *Indicators of Business Expansion and Contractions*. New York: Columbia University Press.

Moravcsik, M.J. (1984). Life in a multidimensional world. *Scientometrics*, *6*(2), 75-86.

Morillo, F., Bordons, M., & Gómez, I. (2001). An approach to interdisciplinarity through bibliometric indicators. *Scientometrics, 51*(1), 203–222.

Morillo, F., Santabárbara, I., & Aparicio, J. (2013). The automatic normalisation challenge: detailed addresses identification. *Scientometrics*, *95*(3), 953-966. doi: 10.1007/s11192-013-0965-0

Müller, R. (2012). Collaborating in life science research groups: the question of authorship. *Higher Education Policy*, *25*(3), 289-311.

Narin, F., Pinski, G., & Gee, H.H. (1976). Structure of the biomedical literature. *JASIS, 27*(1), 25-44.

Newman, M. E. (2004). Who is the best connected scientist? A study of scientific coauthorship networks. *Lecture Notes in Physics*, *650*, 337-370.

Nagpaul, P.S. (1995). Contribution of Indian universities to the mainstream scientific literature: a bibliometric assessment. *Scientometrics*, *32*(1), 11-36.

Narin, F. (1976). *Evaluative bibliometrics: the use of publications and citation analysis in the evaluation of scientific activity*. Washington D.C.: National Science Foundation.

Noyons, E. (1999). *Bibliometric mapping as a science policy and research management tool*. PhD thesis, University of Leiden.

Noyons, E. (2001). Bibliometric mapping of science in a science policy context. *Scientometrics*, *50*(1), 83-98.

Okubo, Y., Miquel, J.F., Frigoletto, L., & Doré, J.C. (1992). Structure of international collaboration in science: Typology of countries through multivariate techniques using a link indicator. *Scientometrics*, *25*(2), 321-351.

Opthof, T., & Leydesdorff, L. (2010). Caveats for the journal and field normalizations in the CWTS ("Leiden") evaluations of research performance. *Journal of Informetrics*, *4*(3), 423–430.

Pao, M.L. (1992). Global and local collaborators: a study of scientific collaboration. *Information Processing & Management*, *28*(1), 99-109.

Park, H.W., & Kang, J. (2009). Patterns of scientific and technological knowledge flows based on scientific papers and patents. *Scientometrics*, *81*(3), 811-820.

Patel, N. (1973). Collaboration in the professional growth of American sociology. *Social Science Information*, *12*(6), 77–92.

Pereira, J.C., & Escuder, M.M.L. (1999). The scenario of Brazilian health sciences in the period of 1981 to 1995. *Scientometrics*, *45*(1), 95-105.

Persson, O., Glanzel, W., & Danell, R. (2004) Inflationary bibliometric values: The role of scientific collaboration and the need for relative indicators in evaluative studies. *Scientometrics*, *60*(3), 421-432.

Polanco, X., François, C., & Keim, J.P. (1998). Artificial neural network technology for the classification and cartography of scientific and technical information. *Scientometrics*, *41*(1), 69-82.

Porter, A.L., & Chubin, D.E. (1985). An indicator of cross-disciplinary research. *Scientometrics*, *8*, 161-176.

Porter, A.L., & Rafols, I. (2009). Is science becoming more interdisciplinary? Measuring and mapping six research fields over time. *Scientometrics*, *81*(3), 719-745.

Pratt, A.D. (1977). A measure of class concentration in bibliometrics. *Journal of the American Society for Information Science*, *28*(5), 285-292.

Price, D.J. de S. (1965). Networks of scientific papers. *Science*, 149, 510–515.

Price, D.J. de S. (1976). A general theory of bibliometric and other cumulative disadvantage processes. *Journal of American Society of Information Science*, *27*(5-6), 292-306.

Price, D.J. de S., & Beaver, D. deB. (1966). Collaboration in an invisible college. *American Psychologist*, *21*(11), 1011-1018.

Pritchard, A. (1969) Statistical bibliography or bibliometrics? *Journal of Documentation*, *25*, 348-349.

Rafols, I., & Meyer, M. (2010). Diversity and network coherence as indicators of interdisciplinarity: case studies in bionanoscience. *Scientometrics*, *82*(2), 263-287.

Ramesh-Babu, A.R., & Singh, Y.P. (1998). Determinants of research productivity. *Scientometrics*, *43*(3), 309-329.

Reagans, R., & Zuckerman, E.W. (2001). Diversity and productivity: the social capital of corporate R&D teams. *Organization Science*, *12*(4), 502–517.

Ruiz-Castillo, J., & Costas, R. (2014). The skewness of scientific productivity. *Journal of Informetrics*, *8*, 917-934.

Safón, V. (2013). What do global university rankings really measure? The search for the X factor and the X entity. *Scientometrics*, *97*(2), 223-244.

Sammon, J.W. (1969): A nonlinear mapping for data structure analysis. *IEEE Transactions on Computers, 18*(5)*, 401–409*.

Sanz-Menéndez, L., & Santesmases, M.J. (1996). Ciencia y política: interacciones entre el Estado y el sistema de investigación. *Zona Abierta*, *75-76*, 1-20.

Schneider, J.W., Larsen, B., & Ingwersen, P. (2009). A comparative study of first and all-author co-citation counting, and two different matrix generation approaches applied for author co-citation analyses. *Scientometrics*, *80*(1), 103-130.

Schubert, A., & Glänzel, W. (1983). Statistical reliability of comparisons based on the citation impact of scientometric publications. *Scientometrics*, *5*, 59-74.

Schubert, A., Glänzel, W., & Braun, T. (1983). Relative citation rate: a new indicator for measuring the impact of publications. In D. Tomov & L. Dimitrova (Eds.). *Proceedings of the First National Conference with International Participation on Scientometrics and Linguistic of the Scientific Text*, pp. 80-81. Varna: Bulgarian Sociological Association.

Seglen, P.O. (1992). The skewness of science. *Journal of the American Society for Information Science*, *43*(9), 628-638.

Small, H. (1973). Co-citation in the scientific literature: a new measure of the relationship between publications. *Journal of the American Society for Information Science*, *24*, 265-269.

Small, H., & Sweeney, E. (1985). Clustering the science citation index® using co-citations. *Scientometrics*, *7*(3-6), 391-409.

Small, H., Sweeney, E., & Greenlee, E. (1985). Clustering the Science Citation Index using co-citations. II. Mapping science. *Scientometrics*, *8*(5), 321-340.

Sonnenwald, D.H. (2007). Scientific collaboration. *Annual Review of Information Science and Technology*, *41*, 643-681.

Subramanyam, K. (1983). Bibliometric studies of research collaboration: a review. *Journal of Information Science*, *6*, 33-38.

Tscharntke, T., Hochberg, M.E., Rand, T.A., Resh, V.H., & Kruass, J. (2007). Author sequence and credit for contributions in multiauthored publications. *Plos Biology*, *5*(1), 13-14.

Testa, J. (2011). *The globalization of Web of Science: 2005-2010*. Thomson Reuters.

Tijssen, R.J.W., de Leeuw, J., & van Raan, A.F. (1987). Quasi-correspondence analysis on scientometric transaction matrices. *Scientometrics*, *11*(5-6), 351-366.

Tijssen, R.J.W., & de Leeuw, J. (1988). Multivariate data-analysis methods in bibliometric studies of science and technology. *Handbook of Quantitative Studies of Science and Technology, North-Holland, Amsterdam*, 705-740.

Todeschini, R. (2011). The j-index: a new bibliometric index and multivariate comparisons between other common indices. *Scientometrics*, *87*(3), 621-639.

Toivanen, H., & Ponomariov, B. (2011). African regional innovation systems: bibliometric analysis of research collaboration patterns. *Scientometrics*, *88*(2), 471–493.

Van den Besselaar, P., Hemlin, S., & van der Weijden, I. (2012). Collaboration and competition in research. *Higher Education Policy*, *25(*3), 263-266.

Van Eck, N.J., & Waltman, L. (2007). VOS: A new method for visualizing similarities between objects. In: R. Decker, and H.J. Lens (Eds.), *Advances in Data Analysis* (pp. 299-306). Berlin Heidelberg: Springer.

Van Eck, N.J., & Waltman, L. (2010). Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics*, *84*(2), 523-538.

Van Leeuwen, T.N., Visser, M.S., Moed, H.F., Nederhof, T.J., & van Raan, A.F.J. (2003). The Holy Grail of science policy: Exploring and combining bibliometric tools in search of scientific excellence. *Scientometrics*, *57*(2), 257-280.

Van Raan, A.F.J. (2004) Measuring science: capita selecta of current main issues. In: H.F. Moed, W. Glänzel, and U. Schmoch (Eds.)*, Handbook of Quantitative Science and Technology Research* (pp. 19-50). Dordrecht: Kluwer Academic Publishers.

Vinkler, P. (1988). An attempt of surveying and classifying bibliometric indicators for scientometric purposes . *Scientometrics*, *13*(5-6), 239-259.

Vinkler, P. (1998). General performance indexes calculated for research institutes of the Hungarian Academy of Sciences based on scientometric indicators. *Scientometrics*, *41*, 185-200.

Vinkler, P. (2010). *The evaluation of research by scientometric indicators*. Cambridge: Chandos Publishing.

Waaijer, C.J., van Bochove, C.A., & van Eck, N.J. (2011). On the map: Nature and Science editorials. *Scientometrics*, *86*(1), 99-112.

Waltman, L. (2012). An empirical analysis of the use of alphabetical authorship in scientific publishing. *Journal of Informetrics*, *6*(4), 700-711. doi: 10.1016/j.joi.2012.07.008

Waltman, L., Tijssen, R.J.W., & van Eck, N.J. (2011). Globalisation of science in kilometres. *Journal of Informetrics*, *5*(4), 574-582. doi: 10.1016/j.joi.2011.05.003

Waltman, L., van Eck, N.J., van Leeuwen, T.N., Visser, M.S., & van Raan, A.F.J. (2011). Towards a new crown indicator: Some theoretical considerations. *Journal of Informetrics*, *5*(1), 37-47. doi: 10.1007/s11192-011-0354-5

Waltman, L., & van Eck, N.J. (2013). Source normalized indicators of citation impact: an overview of different approaches and an empirical comparison. *Scientometrics*, *96*, 699-716. doi: 10.1007/s11192-012-0913-4

Wang, W., Wu, Y., & Pan, Y. (2014). An investigation of collaborations between top Chinese universities: a new quantitative approach. *Scientometrics*, *98*, 1535-1545. doi: 10.1007/s11192-013-1072-y

Ward, J.H. (1963). Hierarchical grouping to optimize an objective function. *Journal of American Statistical Association*, *58*(301), 236-244.

Wasserman, S., & Faust, K., (1994). *Social network analysis: Methods and applications*. Cambridge: Cambridge University Press.

Wildgaard, L., Schneider, J.W., & Larsen, B. (2014). A review of the characteristics of 108 author-level bibliometric indicators. *Scientometrics*, *101*(1), 125–158. doi: 10.1007/s11192-014-1423-3

Wouters, P. (1999). *The citation culture*. PhD thesis. University of Amsterdam.

Wren, J.D., Kozak, K.Z., Johnson, K.R., Deakyne, S.J., Schilling, L.M., & Dellavalle, R.P. (2007). The write position. A survey of perceived contributions to papers based on byline position and number of authors. *EMBO Reports*, *8*(11), 988–991. doi: 10.1038/sj.embor.7401095

Zhou, F., Guo, H.C., Ho, Y.S., & Wu, C.Z. (2007). Scientometric analysis of geostatistics using multivariate methods. *Scientometrics*, *73*(3), 265-279.

Zuccala, A. (2006). Modeling the invisible college. *Journal of the American Society for Information Science and Technology*, *57*(2), 152-168.

Zulueta, M.A., Cabrero, A., & Bordons, M. (1999). Identificación y estudio de grupos de investigación a través de indicadores bibliométricos. *Revista Española de Documentación Científica*, *22*(3), 333-347.

PARTE 2

ARTÍCULOS PUBLICADOS

5. Resumen de las publicaciones

- ✓ **Aportación 1. HJ-Biplot como herramienta de inspección de matrices de datos bibliométricos.**

  Los estudios bibliométricos incluyen indicadores cada vez más sofisticados, un creciente uso de técnicas estadísticas y el desarrollo de nuevas técnicas de visualización de la información. Aunque en los últimos años ha sido especialmente importante el auge que han experimentado las técnicas de visualización, muy por encima del observado para las técnicas de análisis multivariante, estas últimas constituyen una herramienta con un importante potencial en bibliometría.

  Este trabajo pretende poner de manifiesto el interés de los métodos Biplot en los estudios cuantitativos sobre ciencia y tecnología, y exponer sus características frente a otras técnicas de representación simultánea de datos multidimensionales como el Análisis de Correspondencias. Como objeto de estudio se analiza la producción científica en *Web of Science* de los centros propios y mixtos del CSIC en el período 2006-2009 en relación a una serie de indicadores bibliométricos de impacto (porcentaje de artículos Q1, posición normalizada, citas relativas al mundo) y colaboración científica (porcentaje de artículos sin colaboración, porcentaje de artículos en colaboración nacional, porcentaje de artículos en colaboración internacional). Se emplea el HJ-Biplot que proporciona una representación gráfica multivariante en la que los marcadores fila (centros) y marcadores columna (indicadores) pueden ser superpuestos en un mismo sistema de referencia con máxima calidad de representación. Esta técnica permite analizar la relación entre las variables estudiadas, identificar centros que tienen un comportamiento similar en relación a dichas variables y explorar el grado de homogeneidad de las áreas científico-tecnológicas del CSIC en lo que al comportamiento de sus centros se refiere; todo ello a través de un tratamiento y una representación simultánea de variables y centros en un subespacio vectorial de baja dimensión.

  Los resultados obtenidos revelan una asociación positiva entre los indicadores de impacto utilizados y entre el porcentaje de colaboración internacional y las citas recibidas. Las características analizadas son, en parte, dependientes de cada área, ya que los centros e institutos de cada área tienden a posicionarse en la misma zona de los gráficos factoriales resultantes. Sin embargo, existe también cierta heterogeneidad intra-área, de forma que humanidades y ciencias sociales, y alimentos muestran la mayor homogeneidad, mientras que físicas y agrarias presentan mayor variedad de comportamientos entre sus centros e institutos. Por

otro lado, es posible distinguir centros con un comportamiento sobresaliente o singular en el contexto de su área.

Se concluye señalando el potencial e interés del HJ-Biplot en el análisis de datos bibliométricos en la etapa descriptiva. Esta técnica, frente al Análisis de Correspondencias (una potente técnica de ordenación pensada para tablas de contingencia aunque aplicable a matrices de datos en las que tenga sentido trabajar con perfiles), presenta como principales ventajas un ámbito de aplicación mucho más general al ser aplicable a cualquier matriz de datos.

- ✓ **Aportación 2. Unravelling the performance of individual scholars: use of Canonical Biplot analysis to explore the performance of scientists by academic rank and scientific field.**

Existe un creciente interés por los estudios bibliométricos a nivel individual, en los que es importante tener en cuenta no sólo las dimensiones de la actividad científica (colaboración, impacto, nivel de investigación, interdisciplinariedad), sino también factores personales y académicos que pueden influir sobre el desempeño científico. En este contexto, el uso de las técnicas de análisis multivariante puede resultar especialmente relevante.

Este artículo presenta el Biplot Canónico para explorar las diferencias en el desempeño científico de los investigadores del CSIC en el período 2007-2011, medido a través de 17 indicadores bibliométricos, y clasificados según su área de adscripción (química y ciencia de los materiales) y rango académico (investigadores contratados post-doctorales y tres categorías de científicos permanentes). Las publicaciones fueron descargas de la base de datos *Web of Science* y para su correcta asignación a los investigadores se empleó una aplicación semiautomática que normaliza los nombres de autores. El Biplot Canónico proporciona una representación simultánea de filas (grupos) y columnas (indicadores) en un subespacio de baja dimensión de forma que los grupos de investigadores son separados por el máximo poder discriminante entre ellos. Además, dado que los intervalos de confianza se muestran de forma simultánea en el gráfico factorial, se estudia si existen diferencias significativas entre los grupos de investigadores. Como se realizan comparaciones múltiples y el intervalo de confianza está basado en una distribución *t* de Student, se emplea una corrección de Bonferroni que reduce el error tipo I (rechazo indebido de la hipótesis nula). Además, debido a que los test de hipótesis son sensibles a diferentes tamaños muestrales se calculan los tamaños del efecto para algunos pares de grupos a través de la *g* de Hedges.

Se han encontrado dos gradientes, un primer gradiente según el cual al aumentar la categoría científica se incrementa la edad, producción, número de colaboradores, número de artículos altamente citados, así como el porcentaje de documento firmados en último lugar; y un segundo gradiente, en el que se pone de manifiesto las diferencias por áreas. Los tamaños del efecto calculados apoyan la consistencia de los resultados.

Se concluye que el Biplot Canónico es una poderosa herramienta exploratoria para abordar la intrincada estructura de relaciones entre los indicadores bibliométricos y las características individuales de los investigadores; y ofrece algunas ventajas interesantes sobre otras técnicas multivariantes con propósitos similares. Si se hubiese aplicado un MANOVA habría que haber examinado muchas tablas y no se habría obtenido una representación en baja dimensión para la inspección de la estructura subyacente de la matriz de datos. Si se hubiese utilizado un Análisis Discriminante, se habría obtenido un gráfico en dimensión reducida describiendo la estructura de los grupos pero no se habría tenido información directa sobre las variables responsables de la separación entre los grupos y sus correlaciones. Por último, la utilización de tamaños del efecto en los estudios a nivel individual puede resultar de interés pues no siempre es posible obtener tamaños muestrales adecuados para las diferentes categorías o agrupaciones.

- ✓ **Aportación 3. The relationship between the research performance of scientists and their position in co-authorship networks in three fields.**

Las redes de investigación juegan un papel determinante en la producción de nuevo conocimiento, y contribuyen a determinar la estructura cognitiva y social de los campos científicos. Por ello, el análisis de redes sociales ha emergido en tiempos recientes como una aproximación especialmente interesante para el estudio de la co-autoría científica al permitir indagar en las dinámicas de producción de conocimiento propias de cada disciplina, vincular ciertas prácticas a un mejor rendimiento e identificar autores que ocupan posiciones estratégicas dentro de las redes.

Esta investigación profundiza en la estructura de las redes de co-autoría en tres disciplinas (nanociencia y nanotecnología, farmacología y farmacia, y estadística y probabilidad) en España durante el período 2006-2008 a través de la producción científica recogida en *Web of Science*, y explora la relación entre el desempeño científico de los investigadores y la posición de los autores en la redes de co-autoría. Se emplean indicadores de actividad científica (número de artículos, citas, índice-g) y distintas medidas obtenidas mediante análisis de redes sociales (*Pajek*)

relativas a la centralidad (grado de centralidad, centralidad de cercanía, centralidad de intermediación, centralización, centralidad de eigenvector) y cohesión (fuerza de los vínculos, *constraint*, coeficiente de agrupamiento) de la red.

A nivel macro, las redes de farmacología y farmacia, y nanociencia y nanotecnología presentan una estructura similar, más densa que la de estadística y probabilidad, la cual, está más fragmentada y menos conectada (alta *constraint* y menor tendencia a formar cliqués). Muestra de la mayor conexión es también que el componente principal incluye dos terceras partes de los autores en farmacología y farmacia, y nanociencia y nanotecnología, frente a sólo un 28% en estadística y probabilidad.

A nivel micro, se emplea un modelo de regresión de Poisson para estudiar la relación entre desempeño científico, medido a través del índice-g, y los indicadores de redes observándose que existe una relación entre ambos, que varía según el campo y es más débil en estadística y probabilidad. Se observa que el número de co-autores (grado de centralidad) y la fuerza de los vínculos tienen una influencia positiva sobre el índice-g en las tres disciplinas. La cohesión a nivel local influye de forma negativa en dos de los tres campos, donde redes más abiertas y con mayor diversidad de contactos parecen ser más beneficiosas. No se han encontrado ventajas estratégicas claras para los autores que desempeñan un papel de intermediarios (alta intermediación) o que están bien conectados con otros autores (altos valores de *eigenvector*). En términos del índice-g, la posición de los autores en las redes tiene mayor repercusión en las disciplinas con redes de co-autoría más densamente pobladas, característica de las disciplinas de ciencias experimentales analizadas, (farmacología y farmacia, y nanociencia y nanotecnología) que en la disciplina teórica considerada (estadística y probabilidad).

- ✓ **Aportación 4. Acknowledgments in scientific publications: presence in Spanish science and text patterns across disciplines.**

Los agradecimientos que aparecen en los artículos científicos se han convertido en un elemento característico de la comunicación académica y son empleados para reconocer algunas contribuciones relevantes para el desarrollo de una investigación, pero que no alcanzan el estatus de co-autoría científica. De hecho, en la literatura se ha señalado su importancia social, cognitiva e instrumental, así como su potencial como fuente de información sobre colaboración científica (sub-autoría científica). Hasta muy recientemente era muy complicado llevar a cabo estudios acerca de los agradecimientos debido a que esta información no estaba

presente en las bases de datos bibliográficas. Sin embargo, desde el año 2008 la *Web of Science* ha comenzado a incluir esta información lo que abre nuevas posibilidades para su explotación y análisis.

Este trabajo tiene por objeto incrementar nuestro conocimiento acerca de la presencia de los agradecimientos en las publicaciones científicas y explorar su utilidad como fuente de información sobre colaboración científica. En primer lugar, se analiza la presencia de los agradecimientos, con especial énfasis en las diferencias entre áreas, en 38.257 artículos publicados por investigadores españoles en lengua inglesa en el año 2010 en la base de datos *Web of Science*. En segundo lugar, y teniendo en cuenta que la base de datos únicamente incluye los agradecimientos cuando en la sección se realiza alguna mención a la financiación, se introduce una metodología que combina minería de texto y Análisis de Correspondencias para descubrir patrones textuales por disciplina en el campo agradecimientos.

Los agradecimientos están presentes en dos tercios de la investigación española con diferencias significativas entre las áreas (menor frecuencia en ciencias sociales y humanidades) y mayor presencia en revistas de alto impacto. Además, los artículos con agradecimientos muestran un mayor número de autores -que se asocia a una mayor complejidad de la investigación- y una orientación más básica. La metodología introducida revela patrones textuales específicos en las disciplinas seleccionadas (corazón y sistema cardiovascular, economía, biología de la evolución, y estadística y probabilidad). La comunicación interactiva entre pares predomina en los campos de orientación teórica o social (estadística y probabilidad, economía), mientras que el reconocimiento por la asistencia técnica recibida es más común en la investigación experimental (biología de la evolución), y la mención de posibles conflictos de interés emerge de forma destacada en el ámbito clínico (corazón y sistema cardiovascular). Se observa que el contenido de la sección de agradecimientos es heterogéneo y varía según las disciplinas.

El análisis de esta sección emerge como una opción interesante para profundizar en las prácticas de colaboración, debido a que una parte considerable de las mismas quedan fuera del alcance de los indicadores bibliométricos tradicionalmente usados para medir la colaboración en la ciencia (co-autoría). No obstante, sería necesario la inclusión de esta información: a) de todas las revistas (no sólo las escritas en inglés), b) de todos los artículos (no sólo cuando se agradece la financiación), y c) de forma estructurada. La inclusión sistemática de información estructurada de los agradecimientos en las bases de datos facilitaría su procesamiento automático, y tendría un impacto positivo en los estudios de colaboración científica.

6. HJ-Biplot como herramienta de inspección de matrices de datos bibliométricos

Artículo publicado en la Revista Española de Documentación Científica 36(1):e001. doi: http://dx.doi.org/10.3989/redc.2013.1.988. Autores: Adrián A. Díaz-Faes, Borja González-Albo, Mª Purificación Galindo, & María Bordons.

Resumen: El objetivo de este trabajo es poner de manifiesto la utilidad del HJ-Biplot en los estudios bibliométricos. Una representación intuitiva y sencilla, similar a un diagrama de dispersión, pero que captura las estructuras de covariación multivariantes entre los indicadores bibliométricos. Su interpretación no requiere conocimientos estadísticos especializados, basta con saber interpretar la longitud de un vector, el ángulo entre dos vectores y la distancia entre dos puntos. Con este fin, se analiza la actividad científica de los centros propios y mixtos del CSIC durante el período 2006-2009 mediante una serie de indicadores de colaboración e impacto científico. Utilizando un HJ-Biplot es posible interpretar simultáneamente la posición de los centros, representados por puntos, de los indicadores, representados mediante vectores, y de las relaciones entre ambos, en el plano con mayor capacidad informativa. Esto nos permite analizar la situación de cada centro en el contexto de su área y en el contexto general del CSIC e identificar aquéllos que muestran un comportamiento singular. Se concluye que las áreas de Humanidades y Ciencias Sociales, Biología y Biomedicina, Materiales y Químicas son más homogéneas en el comportamiento de sus centros, mientras que Físicas, Agrarias, Recursos Naturales y Alimentos muestran mayor heterogeneidad.

Palabras clave: HJ-Biplot, análisis multivariante, bibliometría, colaboración científica, CSIC.

6.1. Introducción y objetivos

El término “Bibliometría” se atribuye a Pritchard (1969), quien definió el campo como “la aplicación de métodos matemáticos y estadísticos a los libros y otros medios de comunicación”. Años más tarde, Subramanyam (1983) señala que el método bibliométrico facilita el estudio de las relaciones entre las variables a través de la aplicación de técnicas estadísticas como la regresión, correlación o análisis factorial. En las últimas décadas hemos asistido a un importante desarrollo de los estudios bibliométricos, que incluyen indicadores cada vez más sofisticados, un creciente uso de técnicas estadísticas y el desarrollo de nuevas técnicas de visualización de la información. Aunque en los últimos años ha sido especialmente importante el auge

que han experimentado las técnicas de visualización, muy por encima del observado para las técnicas de análisis multivariante, estas últimas también constituyen una interesante herramienta en bibliometría, y a ellas nos vamos a referir en este artículo.
En los estudios bibliométricos abundan representaciones descriptivas uni y bivariantes, siendo las técnicas multivariantes más utilizadas el Análisis de Cluster, el Análisis Factorial con solución en Componentes Principales, y el Análisis de Correspondencias.
El Análisis de Cluster permite clasificar las unidades según similitud, pero no es posible saber qué combinación de variables es la que motiva los agrupamientos que exhibe el correspondiente dendograma. El Análisis Factorial está encaminado a buscar unas pocas variables hipotéticas (conocidas como factores o variables latentes), generadas a partir de las variables observables, que capturen la mayor parte de la información contenida en los datos, pero no proporciona información sobre la similitud entre las unidades objeto de estudio.

La utilización de métodos de representación simultánea de datos multidimensionales se ha visto reducida al Análisis de Correspondencias (Benzécri, 1973), técnica íntimamente relacionada con el Análisis de Componentes Principales, que permite visualizar la posible relación entre un par de variables categóricas, y entre sus respectivas categorías, pero está pensada para trabajar con matrices de frecuencias. En el ámbito bibliométrico esta técnica ha sido empleada por el CNRS para mostrar la evolución de los patrones de publicación a lo largo del tiempo (Miquel y otros, 1995; Doré y otros, 1996; Okubo y otros, 1998; Doré y Ojasoo, 2001), en el análisis jerárquico de la coautoría en las redes de colaboración (Abd el Kader y otros, 1998), así como en el análisis de patentes (Doré y otros, 2000). El Análisis de Correspondencias también ha sido aplicado por Bordons y otros (2004) para estudiar las tendencias en la investigación sobre la aspirina, por Sanz-Casado y Conforti (2005) para analizar la relación entre tipologías documentales y pautas de colaboración científica a nivel micro, por Anuradha y Urs (2007) para identificar patrones de colaboración entre investigadores de la India y por Nagpaul (1995) y Súarez-Balseiro y otros (2009) para evaluar la contribución de los investigadores de las universidades de la India y Puerto Rico, respectivamente, a las publicaciones científicas de mayor impacto internacional.
El HJ-Biplot propuesto originalmente por Galindo (1986), presenta las ventajas del Análisis de Correspondencias, pero es aplicable a cualquier matriz de datos, no solo frecuencias. A pesar de ser una técnica para inspección de matrices de datos multivariantes con menos restricciones que las Correspondencias o el Análisis Factorial, la única referencia de su utilización en el ámbito bibliométrico es Díaz-Faes y otros (2011), donde se aplica el HJ-Biplot para analizar la actividad científica de un conjunto de universidades en el área biosanitaria. No ocurre así en otros campos de la ciencia. Ver, por ejemplo, Cárdenas y otros (2007) que citan aplicaciones en medicina, economía, biología o tecnología ambiental entre otras. Una referencia actual y particularmente interesante es Caballero-Juliá (2011), que presenta el HJ-Biplot como

herramienta en el análisis de grupos de discusión y lo aplica a datos de calidad de vida en ludopatía, en la que puede consultarse exhaustivamente el método.

Para poner de manifiesto la utilidad de la técnica en los estudios bibliométricos, en este estudio se caracteriza la producción científica de los centros propios y mixtos del CSIC en el período 2006-2009 en relación a una serie de indicadores bibliométricos de impacto y colaboración científica. La influencia de la colaboración sobre el impacto de la producción de la investigación de los centros del CSIC ha sido objeto de análisis en un estudio previo (González-Albo y otros, 2012), pero en este caso se presenta una aproximación multivariante a partir de un análisis integrado de indicadores. Se pretende mostrar la utilidad del HJ-Biplot para analizar la relación entre las variables estudiadas, identificar centros que tienen un comportamiento similar en relación a dichas variables y explorar el grado de homogeneidad de las áreas científico-tecnológicas del CSIC en lo que al comportamiento de sus centros se refiere; todo ello a través de un tratamiento y una representación simultánea de variables y centros en un subespacio de baja dimensión.

6.2. Material y métodos

6.2.1. Objeto de estudio

Se ha trabajado con las publicaciones científicas de los centros propios y mixtos del CSIC, recogidas en la base de datos *Web of Science* (WoS), que incluye el *Science Citation Index Expanded* (SCIE), el *Social Sciences Citation Index* (SSCI) y el *Arts & Humanities Citation Index* (AHCI), durante el período 2006-2009. La identificación y codificación de los centros del CSIC se realizó de forma semi-automática (Morillo y otros, 2013) a partir de la producción científica de España – "Spain" en el campo "Address" – descargada de la base de datos WoS en febrero de 2011. Se asignó a cada instituto o centro un código alfa-numérico que permite el posterior tratamiento automático de los datos y caracterizar la actividad científica de los centros propios y mixtos del CSIC con un alto grado fiabilidad (Gómez y otros, 2011a). Asimismo, se realizó una normalización de los títulos de revistas en función de los diferentes campos identificativos de las mismas en WoS – "Full Journal", "Abbreviated Journal", "Serie" e "ISSN" – para su posterior vinculación con los datos de factor de impacto publicados en el *Journal Citation Reports*. El estudio se limita a los ítems citables, que incluyen artículos originales, notas y revisiones. En adelante nos referimos a los ítems citables como artículos. Se ha caracterizado la actividad científica de cada uno de los centros propios y mixtos del CSIC a través de los siguientes indicadores:

a) Indicadores de impacto:

- ✓ *Porcentaje de artículos en el primer cuartil* (Q1), que considera el porcentaje de artículos publicados por cada centro en el 25% de revistas con mayor factor de impacto de cada disciplina (en el caso de revistas asignadas a más de una disciplina se selecciona aquélla en la que ocupa una mejor posición).

- ✓ *Posición normalizada media* (PN), calculada como el cociente entre la posición que ocupa una revista entre las de su disciplina en orden descendente según su factor de impacto y el número total de revistas de la disciplina. Dicho valor se resta de la unidad, de forma que la PN oscila entre 0 y 1. Valores altos de PN indican una buena situación de la revista dentro de su disciplina (Bordons y Barrigón, 1992). Se ha asignado a cada artículo la PN de su revista de publicación, calculándose luego la PN de un centro como el promedio de la PN de todos sus artículos. Se utiliza la PN mejor en el caso de que una revista esté asignada a más de una disciplina. La premisa que subyace al uso del Q1 y la PN es que el factor de impacto de las revistas es un indicador de su prestigio en sus campos de especialización (Moed, 2005). Dado que los valores de factor de impacto varían de forma importante según las disciplinas, se ha preferido utilizar los indicadores Q1 y PN, que no tienen esta limitación y permiten realizar comparaciones entre disciplinas.

- ✓ *Citas relativas al mundo en el período 2006-2009* (CRM), que considera las citas recibidas por los documentos desde el año de publicación hasta 2010 normalizadas respecto a las citas medias recibidas por la producción mundial en cada disciplina. Los datos del total mundial utilizados como referencia proceden de *Thomson Reuters*, que proporcionó una tabla que incluía para cada disciplina las citas medias por artículo recibidas por los artículos de cada año (desde 2005 hasta 2009)[16]. Una explicación detallada sobre la obtención del indicador CRM puede encontrarse en Gómez y otros (2011a). Para el cálculo del indicador citas relativas de cada centro se comparan las citas recibidas por sus artículos con las recibidas por el promedio del mundo, atendiendo a su disciplina y año de publicación. En el caso de revistas asignadas a más de una disciplina en WoS, se han calculado los valores de CRM respecto a las distintas disciplinas, obteniéndose luego el valor medio. Un valor de citas relativas superior a la unidad indica que el centro recibe más citas que el promedio del mundo en sus disciplinas de publicación. El uso de las citas como indicador de la influencia o impacto de la investigación sobre la comunidad científica está ampliamente aceptado en la actualidad, aunque es necesario tener en cuenta sus inconvenientes y limitaciones, repetidamente recogidas en la literatura (Moed, 2005).

[16] Cedidos al MINECO para la convocatoria del Subprograma de Apoyo a Centros y Unidades de Excelencia Severo Ochoa 2011.

b) Indicadores de colaboración:

- ✓ *Porcentaje de artículos sin colaboración* (firmados por un solo centro), *porcentaje de artículos en colaboración nacional* (dos o más centros españoles) y *porcentaje de artículos en colaboración internacional* (al menos un centro extranjero). Los artículos que presentan colaboración nacional e internacional simultáneamente se han considerado en la categoría "internacional".

Hay que señalar que WoS no calcula el factor de impacto para las revistas de Humanidades, por lo que los indicadores Q1 y PN de los centros de estas áreas se refieren sólo a la fracción de su producción que cuenta con dichos indicadores. Por otro lado, se han eliminado del estudio cuatro institutos, adscritos a Humanidades y Ciencias Sociales, por no contar con artículos en revistas con factor de impacto, lo que impedía el cálculo de las variables porcentaje de artículos Q1 y PN: Instituto de Lengua, Literatura y Antropología, Escuela de Estudios Hispanoamericanos, Escuela Española de Historia y Arqueología en Roma e Instituto de Estudios Islámicos y del Oriente Próximo. El conjunto final de datos analizados consistió en una matriz X_{136x6}, 136 centros del CSIC y 6 indicadores bibliométricos.

6.2.2. Métodos Biplot

Los métodos Biplot fueron propuestos por Gabriel (1971) como representaciones gráficas de datos multivariantes, es decir, al igual que un diagrama de dispersión muestra la distribución conjunta de dos variables, un Biplot representa tres o más variables (Gabriel y Odoroff, 1990); son pues, técnicas multivariantes. Usualmente, las filas de la matriz son representadas mediante puntos (marcadores fila) y las columnas con vectores (marcadores columna), siguiendo la terminología introducida por el autor.

Formalmente se definen de la siguiente manera: un Biplot para una matriz de datos X_{nxp}(arreglo rectangular con n filas y p columnas) es una representación gráfica mediante marcadores $g_1, g_2, \ldots, g_n$ para las filas de la matriz de datos X y $h_1, h_2, \ldots, h_p$ para las columnas de X, de forma que el producto escalar $g_i T\, h_j$ aproxime el elemento x_{ij} de la matriz de partida, tan bien como sea posible (Gabriel, 1971). El producto escalar, en el que se basa, es un concepto matemático que en un principio podría suponer una barrera para el usuario, pero su traducción geométrica es sencilla. En este trabajo los datos están contenidos en una matriz X_{136x6} que tiene en filas los 136 centros del CSIC y en columnas, los 6 indicadores bibliométricos. Así, para cada fila i (cada centro del CSIC en nuestro caso) y cada columna j (indicadores bibliométricos) aparece en la matriz de datos un valor x_{ij} que es el valor de ese

marcador j para el centro i. Un Biplot permite representar la fila i de la matriz de datos (un centro del CSIC) con el marcador g_i y la columna j con el vector h_j, de forma que al proyectar el punto g_i sobre el vector h_j, esa proyección coincide con el valor que ese centro ha tenido para ese indicador. Esa es la traducción geométrica del concepto de producto escalar.

El interés práctico reside en que el orden de las proyecciones de cada marcador fila sobre un marcador columna reproduce el orden de la matriz de partida, de forma que analizando la posición de cada unidad (centro) sobre cada variable (indicador), es posible ordenar las unidades en función del valor que toman en ese indicador, y eso puede hacerse para todos y cada uno de los indicadores (ver Figura 6.1).

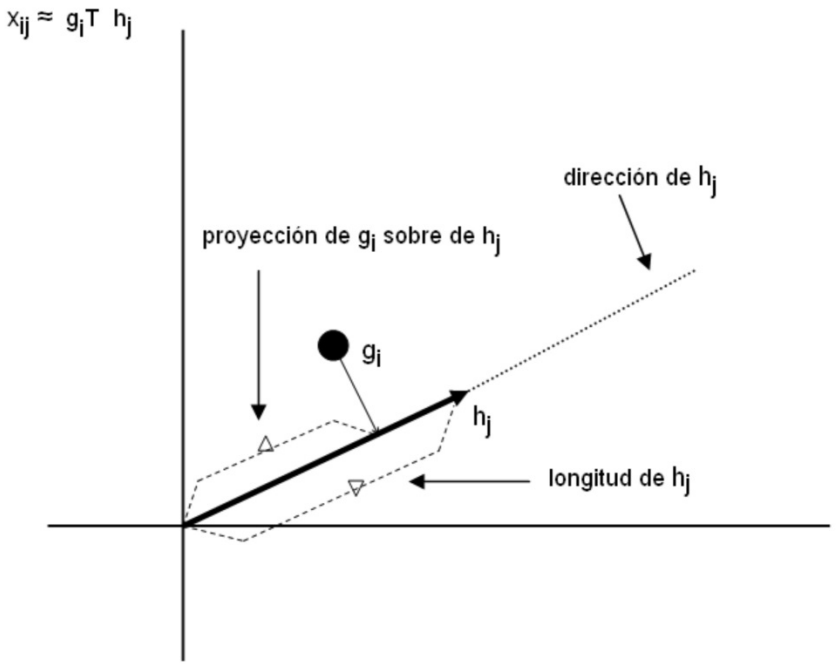


Figura 6.1. Representación geométrica del producto escalar.

Hay infinitas formas de representar un Biplot pero solo algunas tienen propiedades interesantes en el análisis de datos. Los Biplots propuestos originalmente por Gabriel (1971) fueron dos: JK-Biplot, en el cual las filas son representadas con la máxima calidad de representación (fiabilidad de las posiciones de los puntos que representan a los centros) y GH-Biplot en el cual las columnas son representadas con la máxima calidad, pero no las filas.

Galindo (1986) propone el HJ-Biplot como una representación gráfica multivariante de los datos de una matriz X_{nxp}, mediante marcadores $j_1, \ldots, j_n$ para las filas y $h_1, \ldots, h_p$ para las columnas, elegidos de forma que ambos marcadores puedan ser superpuestos en un mismo sistema de referencia con máxima calidad de representación. Al presentar filas y columnas idéntica bondad de ajuste es posible interpretar no sólo la

posición de las filas y de las columnas, sino también las relaciones fila-columna. Los ejes que conforman el sistema de referencia son las Componentes Principales del espacio de los indicadores.

Las reglas para la interpretación del HJ-Biplot son una combinación de las reglas empleadas en otras técnicas como el Escalamiento Multidimensional, el Análisis de Correspondencias, el Análisis Factorial y los Biplots clásicos (Galindo y Cuadras, 1986). En la Figura 6.2 se muestra un ejemplo con cuatro variables y cuatro centros.

- ✓ Las distancias entre los marcadores fila se interpretan como una función inversa de sus similaridades, de tal forma que marcadores próximos (centros del CSIC) son más similares. Esta propiedad permite la identificación de centros con perfiles similares. Cualquier técnica de agrupamiento jerárquico o no jerárquico se puede utilizar para detectar grupos relevantes (Vicente-Tavera, 1992).

- ✓ La longitud de los marcadores columna (vectores) aproximan la desviación típica de los indicadores bibliométricos.

- ✓ Los cosenos de los ángulos entre los vectores columna aproximan las correlaciones entre los indicadores, de modo que ángulos agudos se asocian a indicadores con alta correlación positiva (variables 1 y 2), ángulos obtusos indican correlación negativa (variables 1 y 4) y ángulos rectos señalan variables no correlacionadas (variables 1 y 3). De la misma manera, los cosenos de los ángulos entre los marcadores de los indicadores y los ejes (Componentes Principales) aproximan las correlaciones entre ambos. Para datos estandarizados, las cargas se aproximan a las de los factores en el Análisis Factorial.

- ✓ El orden de las proyecciones ortogonales de los marcadores fila (puntos) sobre un marcador columna (vector) aproxima el orden de los elementos fila (centros) en esa columna (la misma propiedad se cumple para la proyección de los marcadores columna en la dirección definida por un marcador fila). Cuanto mayor es la proyección de un punto sobre un vector, más se desvía el centro de la media de ese indicador bibliométrico. Para una interpretación correcta, la proporción entre las escalas físicas horizontales y verticales ha de ser la misma.

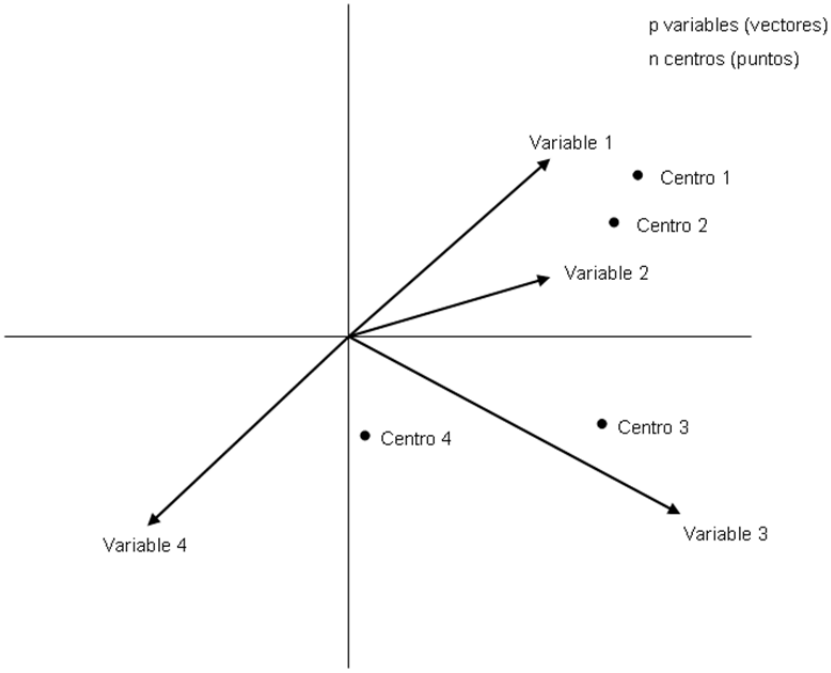


Figura 6.2. Interpretación HJ-Biplot.

Además, como ayuda para una correcta interpretación del gráfico se pueden utilizar algunas medidas adicionales (Galindo y Cuadras, 1986). La correlación al cuadrado entre una variable (indicador) y un factor se interpreta como la contribución relativa del factor al elemento (CRFA). Dado que los factores obtenidos no están correlacionados, la variabilidad de una variable representada en un plano se obtiene sumando las contribuciones de los ejes que conforman el plano, cantidad que se denomina calidad de representación (CLR). Únicamente las variables con CLRs elevadas deben de interpretarse en el plano en cuestión. Una medida análoga de la CLR se puede obtener para cada unidad (centro). Se considera que un centro del CSIC está bien representado cuando se recoge la mayoría de su información (medida a través de la variabilidad) en la dimensión reducida. Debido a que la representación se centra en el origen, la variabilidad de cada centro se mide por su distancia al cuadrado del origen, de modo que la calidad de representación se puede medir por la relación entre la distancia al cuadrado en la dimensión reducida y la distancia al cuadrado en el espacio completo. Geométricamente, es el cuadrado del coseno del ángulo entre el vector en el espacio completo y su proyección sobre el plano de la representación. Si se sumasen las calidades de representación de los ejes factoriales necesarios para absorber toda la variabilidad presente en los datos, la suma de las CLRs sería 1. En este estudio la CLR se valora en una escala de 0 a 1000 puntos.

Conviene resaltar que, a pesar de la aparente similitud entre el HJ-Biplot y el Análisis de Correspondencias, ambos métodos proporcionan resultados diferentes. Ambos métodos pretenden representar las filas y las columnas sobre un subespacio de baja

dimensión en el que sean interpretables sus posiciones relativas, pero las posiciones de los puntos en uno y otro método son diferentes. Las diferencias radican en:

- ✓ La distancia que se define en el hiperespacio: en el Análisis de Correspondencias se trabaja con la distancia Chi-cuadrado y en el HJ-Biplot con la distancia de Mahalanobis. Dos puntos están más cerca, o más lejos, en el hiperespacio, dependiendo de con qué distancia se observen.

- ✓ El peso que se asigna a cada fila/columna: en el Análisis de Correspondencias las líneas de la matriz tienen asignado un peso inversamente proporcional a su total marginal, lo que no ocurre en el HJ-Biplot.

En ambos métodos se trabaja en términos de absorción de inercia, pero inercia es masa por distancia (al cuadrado). Un punto más pesado obviamente viene representado más cerca del origen (si la masa es grande, la distancia tiene que ser corta). Por eso, en un Análisis de Correspondencias las filas (columnas) más frecuentes están cerca del origen y, por tanto, peor representadas. En cambio, en un HJ-Biplot las masas son unitarias, luego inercia es igual a distancia; de forma que a mayor variabilidad, más información, más inercia y más lejos del origen se sitúan los puntos. De esta manera, en las Correspondencias los centros que toman valores más altos, para los diferentes indicadores, son los que vienen peor representados en el gráfico factorial, es decir, serían los centros con menor relevancia en el análisis y sus posiciones en el gráfico factorial podrían ser aparentes. Con el HJ-Biplot sucede justo lo contrario. Los centros propios y mixtos del CSIC en el período 2006-2009 que presentan mayores valores de impacto y colaboración científica son los que tienen más relevancia en el análisis. Obviamente, para nuestro objetivo, el Análisis de Correspondencias no proporciona una solución factorial óptima. También hay que advertir de la gran diferencia existente entre realizar un Análisis de Cluster sobre los datos originales o hacerla, como se hace en este trabajo, sobre las coordenadas del HJ-Biplot. En el primer caso encontraríamos centros con perfiles similares pero no sería posible conocer por qué se han producido esas agrupaciones. En el Análisis de Cluster basado en las coordenadas del Biplot, conocemos además qué indicadores explican las diferentes agrupaciones encontradas.

El análisis se ha llevado a cabo a través del programa MultBiplot desarrollado por Vicente-Villardón (2010) en el entorno de programación orientado a matrices MATLAB. Ejemplos de su aplicación, tanto en su versión integrada en MATLAB como la actual ya compilada, pueden encontrarse en los trabajos de Demey y otros (2008), Vicente-Villardón y otros (2006) o Vicente-Galindo y otros (2011). También existen otras aplicaciones para los métodos Biplot como las desarrolladas en el entorno R por Faria y Demetrio (2011), Nieto-Librero y otros (2011) o Frutos-Bernal y Galindo (2012).

Los datos se han estandarizado por columna debido a las diferentes unidades de medida de las variables. Los centros con CLRs inferiores a 500 puntos no se representan en los gráficos factoriales. Para la selección del tipo de Cluster se han aplicado, con fines exploratorios, Cluster jerárquicos y se ha afinado la solución mediante los métodos no jerárquicos, en concreto, se usó el método K-means y como medida la distancia euclídea. Para la representación factorial se ha tomado el nombre abreviado de los centros y se han clasificado según las ocho áreas científico-técnicas del CSIC (ver Anexo): Humanidades y Ciencias Sociales, Biología y Biomedicina, Recursos Naturales, Ciencias Agrarias, Ciencia y Tecnologías Físicas, Ciencia y Tecnología de Materiales, Ciencia y Tecnología de Alimentos y Ciencia y Tecnologías Químicas (se ha considerado cada centro asignado a su área principal).

6.3. Resultados

La producción científica del CSIC en el periodo 2006-2009 asciende a 28834 artículos. La distribución de la producción por áreas científico-técnicas y el número de centros incluidos en cada área se muestra en la Tabla 6.1.

Tabla 6.1. Número de centros con producción y número de artículos por áreas científico-técnicas del CSIC.

Área CSIC	Nº Centros	Nº Artículos
Humanidades y Ciencias Sociales	19	598
Biología y Biomedicina	24	5345
Recursos Naturales	23	5199
Ciencias Agrarias	14	2448
Ciencia y Tecnologías Físicas	26	5709
Ciencia y Tecnología de Materiales	11	5252
Ciencia y Tecnología de Alimentos	8	1804
Ciencia y Tecnologías Químicas	15	3743
Total	140	28834

* Nota: el sumatorio de artículos es superior al total real porque existe colaboración entre centros de distintas áreas.

Se han retenido tres ejes pues se consigue una inercia acumulada muy elevada, 91,1%, más que suficiente para caracterizar, con garantías, la actividad científica de los centros propios y mixtos del CSIC en relación a las variables de impacto y colaboración consideradas (ver Tabla 6.2).

Tabla 6.2. Valores propios y varianza explicada.

Ejes	Inercia		
	Valor propio	**Var. Explicada**	**Var. Acumulada**
1	20,01	49,43	49,43
2	13,74	23,31	72,74
3	12,20	18,37	91,11
4	7,34	6,66	97,77
5	4,26	2,24	100

Atendiendo a las contribuciones del factor al elemento para las columnas (ver Tabla 6.3), se observa que todas las variables, han de interpretarse en el plano factorial 1-2, a excepción del porcentaje de artículos sin colaboración, que queda mejor recogido en el plano 1-3. La PN, aunque presenta contribuciones ligeramente superiores en el plano 1-3, se analiza en el plano 1-2 por resultar de mayor interés su interpretación junto al resto de indicadores recogidos en dicho plano. En cuanto a las filas, de los 136 centros del CSIC tan sólo seis no han quedado bien recogidos en los tres primeros ejes: Instituto de Ciencia y Tecnología de Polímeros, Instituto de Recursos Naturales y Agrobiología de Sevilla, Instituto de Microbiología Bioquímica, Instituto de Física Fundamental, Instituto de Ciencia y Tecnología de Alimentos y Nutrición y Centro de Investigaciones Físicas Isla de Cartuja[17].

Tabla 6.3. Calidad de representación paras las columnas.

Variables	**Eje 1**	**Eje 2**	**Eje 3**
PN Media	**582**	81	213
% Art. Q1	**775**	85	42
Citas Relativas Mundo	**665**	0	45
% Art. sin colaboración	**446**	15	**535**
% Art. col. nacional	17	**852**	124
% Art. col. internacional	**481**	**366**	143

6.3.1. Análisis del impacto y la colaboración: plano 1-2

En la Figura 6.3 se muestra el gráfico factorial del plano 1-2, donde la inercia acumulada asciende al 72,7%. Los indicadores bibliométricos analizados se representan mediante vectores, mientras que los centros se identifican mediante puntos, cuya etiqueta incluye el nombre abreviado del centro (ver Anexo) y su color varía en función de su área de pertenencia. Los vectores que no están bien

[17] Nótese que Isla de Cartuja es un centro que incluye varios institutos, también visibles en este trabajo. Sólo se asignan al centro aquellos artículos firmados por el centro que no incluyen ninguno de sus institutos.

representados en este plano se presentan de un color más claro (porcentaje de artículos sin colaboración).

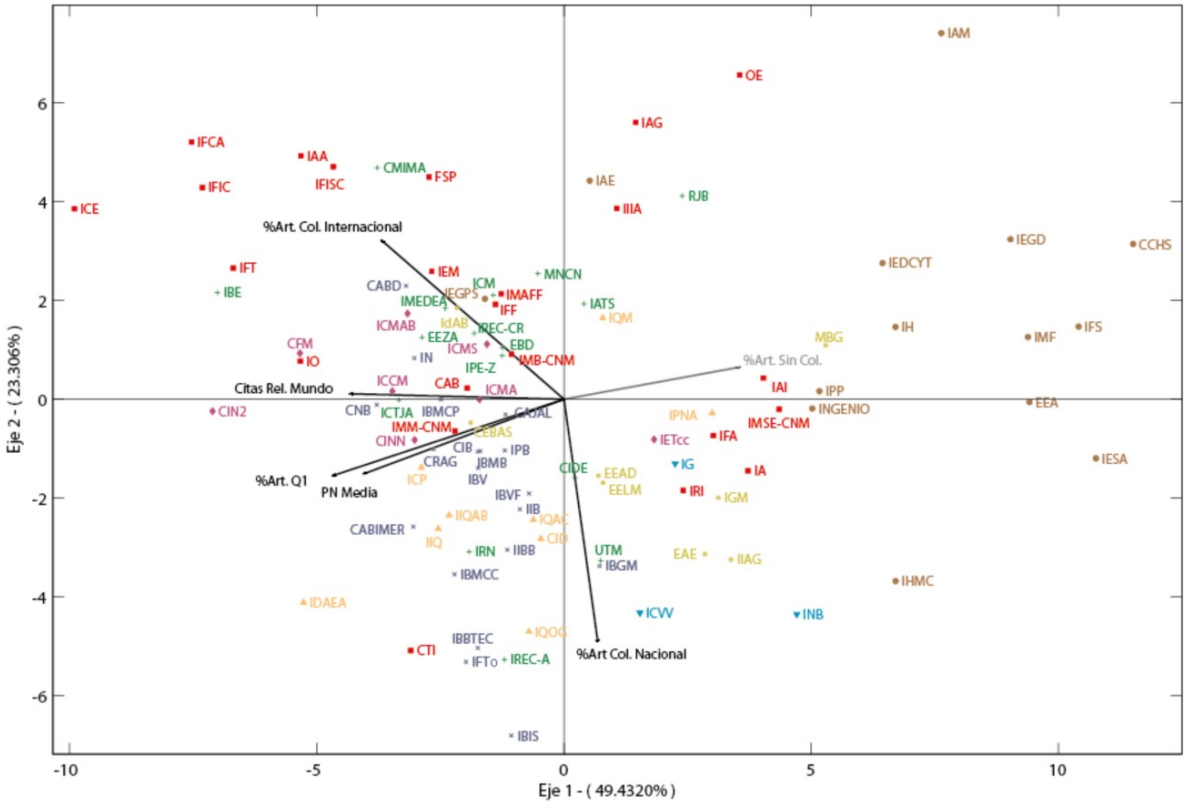


Figura 6.3. Representación factorial resultante del HJ-Biplot, plano 1-2.

Notas: Orden de los cuadrantes: 1º - superior derecho, 2º - superior izquierdo, 3º - inferior izquierdo, 4º - inferior derecho. Se representan los centros con calidades de representación ≥ 500 puntos.

Leyenda con colores de áreas: Humanidades y Ciencias Sociales, Biología y Biomedicina, Recursos Naturales, Agrarias, Físicas, Materiales, Alimentos, Química.

Los indicadores bibliométricos, representados mediante vectores, indican que existe una relación fuerte y directa entre el porcentaje de artículos Q1 en JCR, la PN de la revista en la que se ha publicado y las CRM, siendo esta última variable la más importante para discriminar entre los centros en el eje 1. Igualmente, hay cierta correlación entre el porcentaje de artículos en colaboración internacional y las CRM (ángulo agudo entre los vectores). También cabe señalar que el porcentaje de artículos en colaboración nacional es la variable más importante para discriminar entre individuos en el eje 2, siendo además el indicador que ha presentado una mayor variabilidad en el estudio. Asimismo, esta variable correlaciona inversamente con la producción en colaboración internacional (ángulo casi llano entre ambas variables). Por tanto, los centros situados en la parte izquierda del gráfico factorial publican sus resultados de investigación en revistas de primer nivel y gozan de una visibilidad elevada. Además, se pueden diferenciar en función del alcance de la colaboración: para los centros del segundo cuadrante predomina la asociación con autores adscritos

a centros extranjeros, mientras que la pauta más común para los centros del tercer cuadrante es la coautoría nacional. La variable restante, porcentaje de artículos sin colaboración, y varios centros del área de Alimentos presentan mayores contribuciones en el plano 1-3.

En algunos casos, los centros de una misma área tienden a situarse en la misma zona del gráfico HJ Biplot, lo que indica que presentan características similares en lo que respecta a colaboración e impacto. Es el caso de los centros de Humanidades y Ciencias Sociales (primer cuadrante), Biología y Biomedicina y Químicas (tercer cuadrante) o Alimentos (cuarto cuadrante). Sin embargo, otras áreas son más heterogéneas en lo que respecta al comportamiento de sus centros. Es el caso de Físicas, Ciencias Agrarias o Recursos Naturales, en las que los centros presentan una mayor dispersión en su comportamiento.

En términos generales, atendiendo a la posición de los centros y las variables en el gráfico factorial, se observa que las áreas CSIC que tienden a mostrar un mayor impacto de su producción son Físicas, Químicas y, en menor medida, Biología y Biomedicina, Recursos Naturales y Materiales. Destacan por su alto porcentaje de artículos en revistas Q1 el Instituto de Física Teórica (IFT) (86,5%), el Instituto de Diagnóstico Ambiental y Estudios del Agua (IDAEA) (85,6%) y el Instituto de Biología Evolutiva (IBE) (82,2%). Si se toma como indicador las CRM destacan algunos centros del área de Física como el Instituto de Física Corpuscular (IFIC) o el Instituto de Física de Cantabria (IFCA), junto a algún centro de otras áreas como el Centre d´Investigació en Nanociència i Nanotecnologia (CIN2) del área de Materiales. La PN no discrimina bien entre los centros con valores altos en sus indicadores de impacto y colaboración (el 50% presenta PN comprendidas entre 0,66-0,77), pero es un indicador útil para caracterizar a los centros que han publicado una parte importante de su producción en revistas de bajo factor de impacto dentro de sus respectivas disciplinas, como el Centro de Ciencias Humanas y Sociales[18] (CCHS) (PN=0,41) o el Instituto de Economía, Geografía y Demografía (IEGD) (PN=0,43) adscritos a Humanidades y Ciencias Sociales. Esto es más habitual en centros que publican principalmente en revistas españolas, tal como sucede en Humanidades y Ciencias Sociales, por la peor posición que suelen ocupan estas revistas atendiendo al factor de impacto.

En cuanto a la colaboración, los centros situados en el segundo cuadrante se caracterizan por un elevado número de artículos en colaboración internacional. Entre ellos se puede nombrar el Instituto de Física de Cantabria (IFCA), el Instituto de Astrofísica de Andalucía (IAA), el Centro Mediterráneo de Investigaciones Marinas y Ambientales (CMIMA), el Instituto de Ciencias del Espacio (ICE) y el Instituto de Física Corpuscular (IFIC) adscritos a Físicas (salvo el CMIMA que pertenece al área de

[18] Incluye siete institutos que se analizan de forma independiente en este estudio. Sólo se asignan al CCHS los artículos firmados únicamente por el centro sin incluir ninguno de sus institutos.

Recursos Naturales), que han publicado más del 80% de su producción en coautoría con autores adscritos a centros extranjeros. Resulta llamativa la presencia en esta zona del gráfico de dos centros de Humanidades y Ciencias Sociales, el Instituto de Análisis Económico (IAE) y el Instituto de Estudios Gallegos Padre Sarmiento (IEGPS), que presentan una alta colaboración internacional, 66,7% y 63,6% respectivamente, poco habitual en el área, ya que los centros restantes se sitúan en la parte derecha del gráfico con altas tasas de documentos sin colaboración. Otras áreas como Ciencias Agrarias o Químicas se caracterizan por una mayor investigación en colaboración nacional. En el caso de Biología y Biomedicina coexisten centros con una alta actividad en colaboración nacional, como el Instituto de Biomedicina de Sevilla (IBIS), tercer cuadrante de la Figura 6.3, que ha publicado el 65% de sus artículos en colaboración nacional; y centros con alta actividad en colaboración internacional, como el Centro Andaluz de Biología del Desarrollo (CABD) o el Instituto de Neurociencias de Alicante (IN), segundo cuadrante, que han publicado en coautoría internacional más del 50% de sus documentos. Estos centros se sitúan en la parte izquierda del gráfico factorial porque el impacto de su producción tiende a ser alto. Por el contrario, el Instituto de Nutrición y Bromatología (INB), de Alimentos, y el Instituto de Historia de la Medicina y de la Ciencia López Piñero (IHMC), de Humanidades y Ciencias Sociales, se sitúan en la parte inferior del cuarto cuadrante pues publican, principalmente, en colaboración nacional pero su impacto en la comunidad científica es menor.

6.3.2. Análisis del impacto y la colaboración: plano 1-3

La absorción de inercia en el plano factorial 1-3 es del 67,8% (ver Figura 6.4). Esta representación resulta de interés para caracterizar el porcentaje de artículos sin colaboración, al ser óptima la calidad de representación para este indicador.

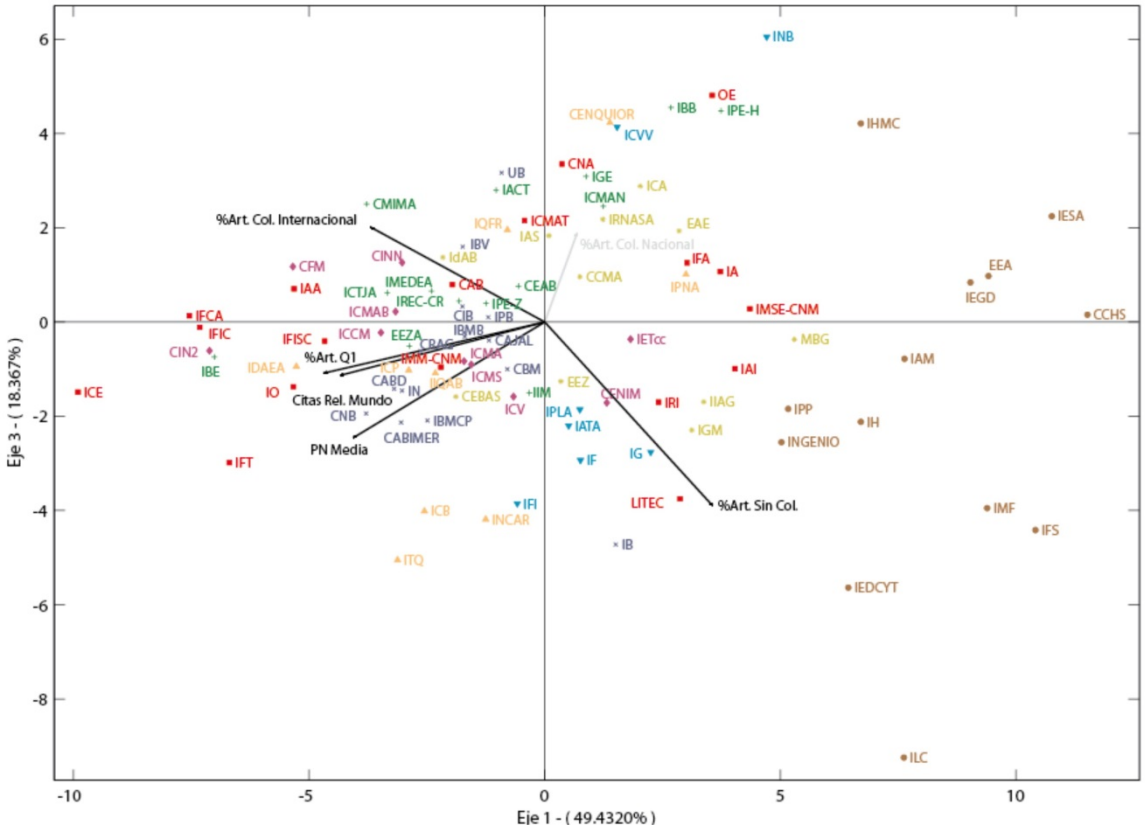


Figura 6.4. Representación factorial resultante del HJ-Biplot, plano 1-3.

Notas: Orden de los cuadrantes: 1º - superior derecho, 2º - superior izquierdo, 3º - inferior izquierdo, 4º - inferior derecho. Se representan los centros con calidades de representación ≥ 500 puntos.

Leyenda con colores de áreas: Humanidades y Ciencias Sociales, Biología y Biomedicina, Recursos Naturales, Agrarias, Físicas, Materiales, Alimentos, Química.

Se observa una relación inversa entre el porcentaje de artículos sin colaboración y los realizados en coautoría internacional. Esta variable parece independiente de los indicadores relativos a visibilidad e impacto. Los centros situados en el cuarto cuadrante presentan un patrón de publicación en el que predominan los artículos firmados por un único centro, práctica común entre los centros del área de Humanidades y Ciencias Sociales (Gómez y otros, 2011b). Destacan por su situación algo periférica en el cuarto cuadrante algunos institutos con más de 2/3 partes de su producción firmada por un solo centro, como son el Instituto de Lenguas y Culturas del Mediterráneo y Oriente Próximo (ILC) (87%), el Instituto de Filosofía (IFS) (73,7%), la Institución Milá Fontanals (IMF) (68,4%) y el Instituto de Estudios Documentales sobre Ciencia y Tecnología (IEDCYT) (72,9%). En los institutos más próximos al centro de gravedad comienza a incrementarse la actividad en colaboración y los resultados de investigación se publican en revistas de mayor visibilidad internacional, así el Instituto de Gestión de la Innovación y del Conocimiento (INGENIO), el Instituto de Políticas y Bienes Públicos (IPP) han publicado un 44,2% y un 38,5% de artículos en revistas indexadas en el Q1. Este plano recoge, además, a varios centros del área de Alimentos

que no quedaban bien representados en el plano factorial 1-2. Se observa que la mayor parte de los institutos de esta área muestran alta actividad sin colaboración, situándose en el cuarto cuadrante de la Figura 6.4; a excepción de dos institutos con escasa actividad sin colaboración, pero con alta colaboración nacional, que se sitúan en el primer cuadrante. Finalmente, es interesante señalar la presencia de algunos centros como la Misión Biológica de Galicia (MBG), de Agrarias, o el Instituto de Automática Industrial (IAI), de Físicas, que presentan una alta actividad sin colaboración que los diferencia de otros centros de sus áreas respectivas.

6.3.3. Clusters según tipo de colaboración

A través de las coordenadas Biplot se han calculado los Clusters (método K-means, distancia euclídea). Se observa en el gráfico factorial (Figura 6.5) que los centros forman conglomerados en función de su comportamiento en las variables de impacto y colaboración. Las calidades de representación para cada conglomerado en el plano 1-2 se exponen en la Tabla 6.4.

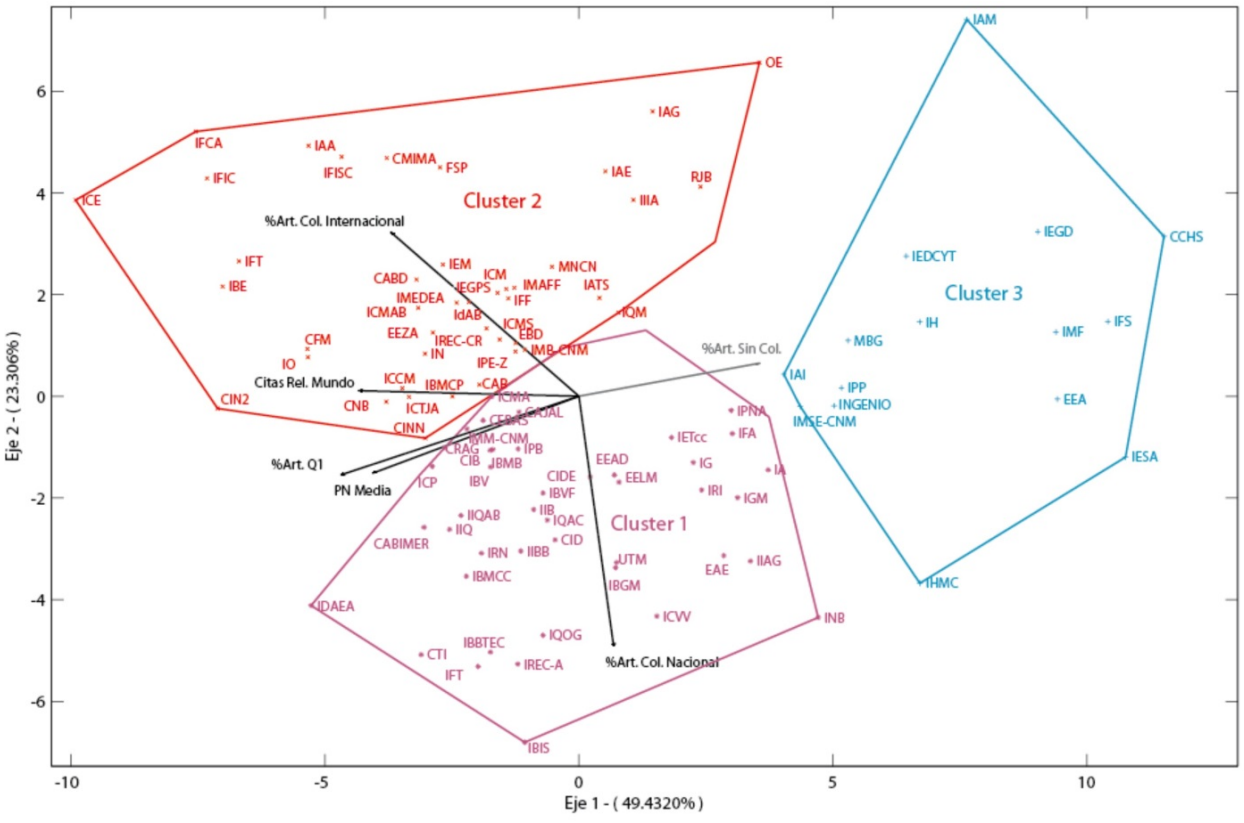


Figura 6.5. Representación factorial del HJ-Biplot por Clusters, plano 1-2.

Notas: se representan los centros con calidades de representación ≥ 500 puntos.

Tabla 6.4. Conglomerados y calidades de representación en el plano 1-2

	n	CLR - 1	CLR - 2	CLR 1-2
Cluster 1	73	0,23	99,24	99,47
Cluster 2	47	53,77	43,82	97,59
Cluster 3	16	93,41	2,48	95,89

La descripción general de los clusters en función de las seis variables utilizadas se muestra en la Tabla 6.5. Observamos importantes diferencias entre clusters en las prácticas de colaboración predominantes en cada caso, de forma que en los centros del Cluster 1 predomina la colaboración de orientación nacional, mientras que la colaboración internacional predomina en el Cluster 2, y la producción sin colaboración entre centros impera en el Cluster 3. La Figura 6.6 muestra la composición de los clusters atendiendo al área científico-técnica de sus centros. Considerando la distribución total de centros por áreas, observamos que el Cluster 1 se caracteriza por una alta presencia relativa de los centros de las áreas de Biología y Biomedicina, Ciencias Agrarias, Alimentos y Químicas; y una ausencia de centros de Humanidades y Ciencia Sociales. Así, en el Cluster 1 predominan los centros de Biología y Biomedicina (26%), Químicas (19%) y Agrarias (16%). Todos los centros e institutos del área de Alimentos quedan integrados en el Cluster 1. En el Cluster 2 predominan los centros adscritos a Físicas (36%), Recursos Naturales (30%) y Materiales (15%), que muestran una alta presencia relativa. En el Cluster 3 predominan las Ciencias Sociales y Humanidades (81%).

Tabla 6.5. Descriptivos de los indicadores de impacto y colaboración según Clusters.

	PN	Q1	CRM	% sin colaboración	% colaboración nacional	% colaboración internacional
Cluster 1	0,73 ± 0,01	59,19 ± 1,44	1,20 ± 0,04	19,97 ± 1,44	43,62 ± 1,31	36,41 ± 1,20
Cluster 2	0,73 ± 0,01	59,09 ± 2,00	1,43 ± 0,08	13,17 ± 1,05	26,02 ± 1,40	60,81 ± 1,78
Cluster 3	0,60 ± 0,03	20,15 ± 3,49	0,60 ± 0,09	51,20 ± 4,37	30,95 ± 4,11	17,85 ± 2,96
Total	0,72 ± 0,01	54,56 ± 1,55	1,21 ± 0,04	21,29 ± 1,39	36,05 ± 1,21	42,66 ± 1,56

Nota: datos expresados como media ± desviación típica.

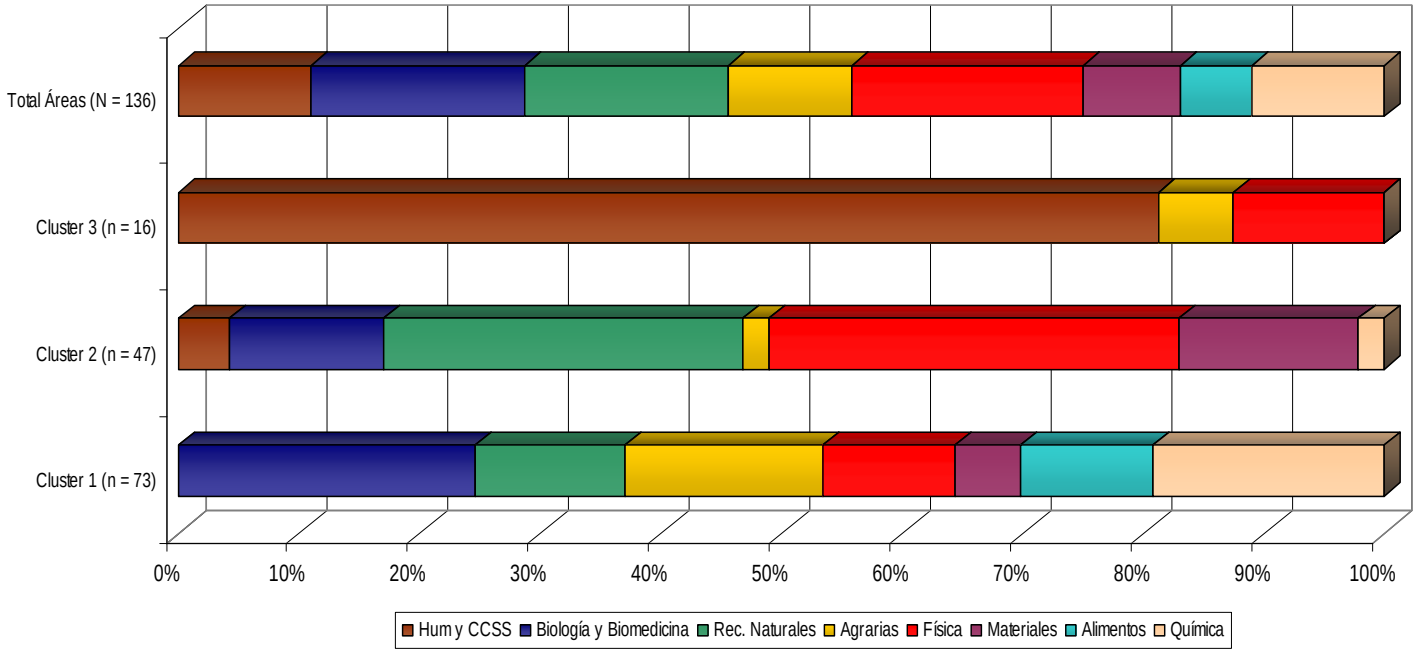


Figura 6.6. Configuración de los conglomerados por áreas.

En promedio, los Clusters 1 y 2 presentan mayores valores de impacto que el Cluster 3. En la Figura 6.5 se identifican los centros con más impacto en la parte izquierda de los Clusters 1 y 2. En el Cluster 3 los centros tienen en promedio menos impacto y se caracterizan por la publicación de artículos firmados por un único centro así como por englobar la mayor parte de los centros de Humanidades y Ciencias Sociales (todos excepto el IAE y el IEGPS antes mencionados). Resulta llamativa la presencia en el Cluster 3 del Instituto de Microelectrónica de Sevilla (IMSE-CNM) y el Instituto de Automática Industrial (IAI), del área de Físicas, así como de la Misión Biológica de Galicia (MBG), del área de Ciencias Agrarias, que se caracterizan por su alta actividad sin colaboración, baja colaboración internacional, y bajo porcentaje de documentos en revistas Q1 -comparado con la media de sus áreas-.

6.4. Discusión y conclusiones

En este estudio se ha analizado la distribución de los institutos y centros del CSIC en función de sus prácticas de colaboración y su impacto mediante la técnica HJ-Biplot que permite realizar una representación gráfica de datos multivariantes en el que filas y columnas pueden ser superpuestas en un mismo sistema de referencia con máxima calidad de representación. Se pone de manifiesto la relación entre variables, observándose una correlación fuerte y directa entre los indicadores basados en citas y factor de impacto, de forma que el número de citas recibidas va asociado a un mayor alcance de la colaboración. Así, el estudio permite observar que las características analizadas son, en parte, dependientes de cada área, ya que los centros e institutos de

cada área tienden a posicionarse en la misma zona de los gráficos factoriales resultantes. Sin embargo, existe también cierta heterogeneidad intra-área, de forma que Humanidades y Ciencias Sociales, y Alimentos muestran la mayor homogeneidad, mientras que Físicas y Agrarias presentan mayor variedad de comportamientos entre sus centros e institutos.

Los resultados obtenidos son consistentes con otros estudios que señalan la influencia positiva de la colaboración internacional sobre el impacto de la investigación (ver por ejemplo, Glänzel y Schubert, 2001). El análisis de los datos ha puesto de manifiesto que las áreas CSIC de un mayor impacto y visibilidad internacional son Físicas, caracterizada por la coautoría internacional; y Químicas, donde la producción con autores adscritos a centros nacionales es el patrón más común. Las mayores tasas de colaboración internacional se dan en las áreas de Físicas y Recursos Naturales. En el caso de Ciencias Sociales y Humanidades, se aprecian diferencias entre los centros de Humanidades, por un lado, con un escaso ratio de colaboración e impacto; y los centros más próximos a Ciencias Sociales, por otro lado, que tienden a presentar mayor colaboración y visibilidad internacional. No obstante, hay que señalar que los resultados aquí expuestos sobre Humanidades sólo representan un pequeño fragmento de la actividad científica del área, pues sus centros publican una parte importante de sus resultados en libros y revistas de ámbito regional no siempre incluidas en WoS y, por otro lado, la base de datos no calcula el factor de impacto para las revistas de Humanidades (González-Albo y otros, 2012).

El estudio actual permite analizar el comportamiento específico de cada centro, y su situación en el contexto de su área y del total de los centros e institutos del CSIC. Así, es posible identificar centros que se salen del patrón general de su área, y determinar qué faceta de su actividad les hace singulares. Exponemos a continuación algunos ejemplos en varias de las áreas analizadas.

En Humanidades y Ciencias Sociales hay que destacar el Instituto de Análisis Económico (IAE) y el Instituto de Estudios Gallegos Padre Sarmiento (IEGPS) que se sitúan en el segundo cuadrante de la Figura 6.3, lejos del resto de institutos de su área ubicados principalmente en el primer cuadrante; lo que se explica por su alta orientación internacional. El IAE participa en diversos proyectos europeos (CSIC, 2009) y publica prácticamente toda su producción en revistas internacionales, mientras que el IEGPS tiene una orientación más regional y su producción en WoS sólo representa 1/3 de su actividad científica (Gómez y otros, 2011a), pero es la que queda reflejada en el estudio actual. También en el área de Humanidades y Ciencias Sociales llama la atención por su distante posición en el cuarto cuadrante el Instituto de Historia de la Medicina y de la Ciencia López Piñero (IHMC), cuya actividad se caracteriza por una alta orientación a la colaboración nacional (68%), pero situando su producción en revistas de poca visibilidad (25% en el primer cuartil).

Los institutos de Recursos Naturales tienden a concentrarse en el segundo cuadrante de la Figura 6.3. No obstante, algunos institutos con un comportamiento "atípico" o singular se sitúan en otras zonas del gráfico. Cabe señalar al Real Jardín Botánico (RJB), situado en el primer cuadrante, con una tasa reseñable de colaboración internacional (59%) pero con poca producción en revistas del primer cuartil (33%). Un análisis detallado de los datos pone en evidencia que sus publicaciones tienden a concentrarse en revistas muy especializadas, sobre todo de botánica, alejadas de las primeras posiciones ocupadas por revistas de ámbito más general. Por otro lado, en el tercer y cuarto cuadrante se sitúan algunos institutos que desarrollan una investigación con un carácter más aplicado, como son el Instituto de Investigación en Recursos Cinegéticos de Albacete (IREC-A) que presenta una alta tasa de colaboración nacional (69%) y una visibilidad notable (65% de artículos en el primer cuartil y una PN de 0,71), y la Unidad de Tecnología Marina (UTM), que presta un servicio de apoyo logístico y técnico (CSIC, 2009), suele colaborar con centros nacionales (61%) y tiene un visibilidad algo menor (47% de artículos en primer cuartil y PN=0,68).

Aunque los institutos de Materiales tienden a concentrarse en el segundo cuadrante de la Figura 6.3, destaca la situación en el cuarto cuadrante del Instituto de Ciencias de la Construcción Eduardo Torroja (IETcc). Este centro se aleja del patrón de su área debido a un menor ratio de colaboración internacional (32%), lo que puede estar asociado al desarrollo de una investigación más aplicada, ya que presta servicios de apoyo científico-técnico al sector de la construcción (CSIC, 2009).

Los institutos de Ciencias Agrarias aparecen bastante dispersos, aunque predominan en la parte derecha de la Figura 6.3. Sin embargo, en la mitad izquierda se identifican dos institutos que sobresalen por su alta actividad en revistas Q1 y colaboración internacional, como son el Instituto de Agrobiotecnología de Navarra (IdAB), ubicado en el segundo cuadrante, y el Centro de Edafología y Biología Aplicada del Segura (CEBAS), en el tercer cuadrante (presenta menor colaboración internacional que el IdAB).

El área de Físicas incluye institutos dispersos por los cuatro cuadrantes de la Figura 6.3. En el segundo cuadrante se incluyen centros orientados a la "Big Science", como el Instituto de Ciencias del Espacio (ICE) o el Instituto de Astrofísica de Andalucía (IAA), que tienen una alta orientación internacional de la investigación, que se publica en revistas de alto factor de impacto. Por el contrario, en el cuarto cuadrante se sitúan centros que trabajan en la línea de tecnologías físicas o informáticas (CSIC, 2009) y cuya investigación tiene una orientación más nacional, como el Instituto de Acústica (IA).

En conclusión, el HJ-Biplot se ha revelado como una herramienta multivariante sumamente útil en el análisis de datos bibliométricos en la etapa descriptiva. Este método, frente al Análisis de Correspondencias -una potente técnica de ordenación

pensada para tablas de contingencia aunque aplicable a matrices de datos en las que tenga sentido trabajar con perfiles-, presenta como principales ventajas un ámbito de aplicación mucho más general al ser aplicable a cualquier matriz de datos y la posibilidad de detectar qué indicadores bibliométricos son los responsables de las agrupaciones de los centros. La aplicación del HJ-Biplot al estudio de la producción científica del CSIC nos ha permitido caracterizar la actividad de las áreas en cuanto a colaboración e impacto se refieren e identificar centros con un comportamiento sobresaliente o singular que los diferencia del resto de su área.

Agradecimientos

Agradecemos los comentarios de Isabel Gómez Caridad sobre una versión previa de este documento. Adrián A. Díaz-Faes cuenta con una beca predoctoral de la Junta de Ampliación de Estudios – Consejo Superior de Investigaciones Científicas (JAE-CSIC). Este artículo ha sido realizado en el marco de los proyectos 200410E605 y CSO2008-06310.

Referencias

Abd el Kader, M., Ojasoo, T., Miquel, J.F., Okubo, Y., & Doré, J.C. (1998). Hierarchical author networks: an analysis of European Molecular Biology Laboratory (EMBL) publications. *Scientometrics*, *42*(3), 405-421.

Anuradha, K.T., & Urs, S.R. (2007). Bibliometric indicators of Indian research collaboration patterns: a correspondence analysis. *Scientometrics*, *71*(2), 179-189.

Benzécri, J.P. (1973). *L'analyse de Données. 2. L'analyse des correspondances*. Paris: Dunod.

Bordons, M., & Barrigón, S. (1992). Bibliometric analysis of publications of Spanish pharmacologists in the SCI (1984-1989) Part I. *Scientometrics*, *25*(3), 425-446.

Bordons, M., Bravo, C., & Barrigón, S. (2004). Time-tracking of the research profile of a drug using bibliometric tools. *Journal of the American Society for Information Science and Technology*, *55*(5), 445-461.

Caballero-Juliá, D. (2011). *El HJ-Biplot como herramienta en el análisis de grupos de discusión. Calidad de vida en la ludopatía: una propuesta sociológica.* (Tesis de maestría). http://hdl.handle.net/10366/108778.

Cárdenas, O., Galindo, M.P., & Vicente-Villardón, J.L. (2007). Los métodos Biplot: evolución y aplicaciones. *Revista Venezolana de Análisis de Coyuntura*, *13*(1), 279-303.

Díaz-Faes, A.A., Benito-García, N., Martín-Rodero, H., & Vicente-Villardón, J.L. (2011). Propuesta de aplicabilidad del método multivariante gráfico Biplot a los estudios bibliométricos en biomedicina. *Actas XIV Jornadas Nacionales de Información y*

Documentación en Ciencias de la Salud, p. 66. Cádiz, España: Biblioteca Virtual del Sistema Sanitario Público de Andalucía.

Doré, J.C., Dutheuil, C., & Miquel, J.F. (2000). Multidimensional analysis of trends in patent activity. *Scientometrics*, *47*(3), 475-492.

Doré, J.C., & Ojasoo, T. (2001). How to analyze publication time trends by correspondence factor analysis: Analysis of publications by 48 countries in 18 disciplines over 12 years. *Journal of the American Society for Information Science and Technology*, *52*(9), 763-769.

Dore, J.C., Ojasoo, T., Okubo, Y., Durand, T., Dudognon, G., & Miquel, J.F. (1996). Correspondence factor analysis of the publication patterns of 48 countries over the period 1981-1992. *Journal of the American Society for Information Science*, *47*(8), 588-602.

Faria, J.C., & Demetrio, C.G.B. (2011). BPCA: Biplot of multivariate data based on Principal Components Analysis [Programa informático]. ESALQ, USP, Brasil. http://cran.r-project.org/web/packages/bpca/citation.html

Frutos-Bernal, E. (2012). GGEBiplotGUI: Interactive GGE Biplots in R [Programa informático]. Salamanca, España: Departamento de Estadística, Universidad de Salamanca. http://cran.r-project.org/web/packages /GGEBiplotGUI/index.html

Gabriel, K.R. (1971). The biplot graphic display of matrices with application to principal component analysis. *Biometrika*, *58*(3), 453-467.

Gabriel, K.R., & Odoroff, C.L. (1990). Biplots in biomedical research. *Statistics in Medicine*, *9*, 469-485.

Galindo, M.P. (1986). Una alternativa de representación simultánea: HJ-Biplot. *Qüestioó*, *10*(1), 13-23.

Galindo, M.P., & Cuadras, C. (1986). *Una extensión del método Biplot y su relación con otras técnicas*. Publicaciones de Bioestadística y Biomatemática, 17. Barcelona: Universidad de Barcelona.

Glanzel, W., & Schubert, A. (2001). Double effort = double impact? A critical view at international co-authorship in chemistry. *Scientometrics*, *50*(2), 199-214.

Gómez, I., Bordons, M., Morillo, F., Moreno, L., Aparicio, J., Díaz-Faes, A.A., & González-Albo, B. (2011a). *La actividad científica del CSIC a través de indicadores bibliométricos (Web of Science, 2006-2010)*. Madrid: IEDCYT, CCHS, CSIC. http://hdl.handle.net/10261/48118.

Gómez, I., Bordons, M., Morillo, F., Moreno, L., & González-Albo, B. (2011b). *La actividad científica del CSIC: Indicadores de producción e impacto por tipo de colaboración (WoS, 2004-2009)*. Madrid: IEDCYT, CCHS, CSIC. http://hdl.handle.net/10261/38113.

González-Albo, B., Moreno, L., Morillo, F., & Bordons, M. (2012). Indicadores bibliométricos para el análisis de la actividad de una institución multidisciplinar: el CSIC. *Revista Española de Documentación Científica*, *35*(1), 9-37.

Miquel, J.F., Ojasoo, T., Okubo, Y., Paul, A., & Doré, J.C. (1995). World science in 18 disciplinary areas: comparative evaluation of the publication patterns of 48 countries over the period 1981–1992. *Scientometrics*, *33*(2), 149–167.

Moed, H.F. (2005). *Citation analysis in research evaluation*. Dordrecht, The Netherlands: Springer.

Morillo, F., Aparicio, J., González-Albo, B., & Moreno, L. (2012). Towards the automation of address identification. *Scientometrics, 94*(1), 207-224. doi: 10.1007/s11192-012-0733

Nagpaul, P.S. (1995). Contribution of Indian Universities to the mainstream scientific literature: a bibliometric assessment. *Scientometrics*, *32*(1), 11-36.

Nieto-Librero, A.N., Baccala, N., & Galindo, M.P. (2011) MultibiplotGUI: Multibiplot Analysis in R [Programa informático]. Salamanca, España: Departamento de Estadística, Universidad de Salamanca. http://cran.r-project.org/web/packages/multibiplotGUI/index.html

Okubo, Y., Doré, J.C., Ojasoo, T., & Miquel, J.F. (1998). A multivariate analysis of publication trends in the 1980s with special reference to South-East Asia. *Scientometrics*, *41*(3), 273-289.

Pritchard, A. (1969) Statistical bibiography or bibliometrics? *Journal of Documentation*, *25*, 348-349.

Sanz-Casado, E., & Conforti, N. (2005). Análisis de la actividad científica de la Facultad de Humanidades de la Universidad de Mar de Plata, durante el periodo 1998-2001. *Revista Española de Documentación Científica*, *28*(2), 196-205.

Suárez-Balseiro, C., García-Zorita, C., & Sanz-Casado, E. (2009). Multi-authorship and its impact on the visibility of research from Puerto Rico. *Information Processing and Management*, *45*, 469-476.

Subramanyam, K. (1983). Bibliometric studies of research collaboration: a review. *Journal of Information Science*, *6*, 33-38.

Vicente-Tavera, S. (1992). *Las técnicas de representación de datos multidimensionales en el estudio del índice de producción industrial en la C.E.E.* (Tesis doctoral no publicada). Universidad de Salamanca, Departamento de Estadística, Salamanca España.

Vicente-Villardón, J.L. (2010). Multbiplot: a packaged for multivariate analysis using Biplots. (versión 1.0) [Programa informático]. Salamanca, España: Departamento de Estadística, Universidad de Salamanca. http://biplot.dep.usal.es/classicalbiplot/.

Anexo. Relación de centros propios y mixtos del CSIC por áreas científico-técnicas (2006-2009)

Centros
Área 1. Humanidades y Ciencias Sociales
Centro de Ciencias Humanas y Sociales (CCHS), Madrid
Escuela de Estudios Árabes (EEA), Granada
Escuela de Estudios Hispanoamericanos (EEHA), Sevilla
Escuela Española de Historia y Arqueología de Roma (EEHAR)
Institución Milá y Fontanals (IMF), Barcelona
Instituto de Análisis Económico (IAE), Barcelona
Instituto de Arqueología (IAM), Mérida
Instituto de Economía, Geografía y Demografía (IEGD), Madrid
Instituto de Estudios Documentales sobre la Ciencia y la Tecnología (IEDCYT), Madrid
Instituto de Estudios Gallegos "Padre Sarmiento" (IEGPS), A Coruña
Instituto de Estudios Islámicos y del Oriente Próximo (IEIOP), Zaragoza
Instituto de Estudios Sociales Avanzados de Andalucía (IESA), Córdoba
Instituto de Filosofía (IFS), Madrid
Instituto de Gestión de la Innovación y del Conocimiento (INGENIO), Valencia
Instituto de Historia (IH), Madrid
Instituto de Historia de la Medicina y de la Ciencia "López Piñero" (IHMC), Valencia
Instituto de Lengua, Literatura y Antropología (ILLA), Madrid
Instituto de Lenguas y Culturas del Mediterráneo y Oriente Próximo (ILC), Madrid
Instituto de Políticas y Bienes Públicos (IPP), Madrid
Área 2. Biología y Biomedicina
Centro Andaluz de Biología del Desarrollo (CABD), Sevilla
Centro Andaluz de Biología Molecular y Medicina Regenerativa (CABIMER), Sevilla
Centro de Biología Molecular "Severo Ochoa" (CBM), Madrid
Centro de Investigación Cardiovascular (CIC), Barcelona
Centro de Investigación en Agrogenómica (CRAG), Barcelona
Centro de Investigaciones Biológicas (CIB), Madrid
Centro Nacional de Biotecnología (CNB), Madrid
Instituto de Biología Molecular de Barcelona (IBMB)
Instituto de Biología Molecular y Celular de Plantas (IBMCP), Valencia
Instituto de Biología Molecular y Celular del Cáncer (IBMCC), Salamanca
Instituto de Biología y Genética Molecular (IBGM), Valladolid
Instituto de Biomedicina de Sevilla (IBIS)
Instituto de Biomedicina de Valencia (IBV)
Instituto de Biomedicina y Biotecnología de Cantabria (IBBTEC), Santander
Instituto de Bioquímica (IB), Madrid
Instituto de Bioquímica Vegetal y Fotosíntesis (IBVF), Sevilla
Instituto de Farmacología y Toxicología (IFTo), Madrid
Instituto de Investigaciones Biomédicas "Alberto Sols" (IIB), Madrid
Instituto de Investigaciones Biomédicas de Barcelona (IIBB)
Instituto de Microbiología Bioquímica (IMB), Salamanca
Instituto de Neurobiología "Ramón y Cajal" (CAJAL), Madrid
Instituto de Neurociencias (IN), Alicante
Instituto de Parasitología y Biomedicina “López-Neyra” (IPB), Granada
Unidad de Biofísica (UB), Leioa [Vizcaya]
Área 3. Recursos Naturales
Centro de Estudios Avanzados de Blanes (CEAB), Girona
Centro de Investigaciones sobre Desertificación (CIDE), Valencia
Centro Mediterráneo de Investigaciones Marinas y Ambientales (CMIMA), Barcelona
Estación Biológica de Doñana (EBD), Sevilla
Estación Experimental de Zonas Áridas (EEZA), Almería
Instituto Andaluz de Ciencias de la Tierra (IACT), Granada

Instituto Botánico de Barcelona (IBB), Barcelona
Instituto de Acuicultura de Torre de la Sal (IATS), Castellón
Instituto de Biología Evolutiva (IBE), Barcelona
Instituto de Ciencias de la Tierra "Jaume Almera" (ICTJA), Barcelona
Instituto de Ciencias del Mar (ICM), Barcelona
Instituto de Ciencias Marinas de Andalucía (ICMAN), Cádiz
Instituto de Geología Económica (IGE), Madrid
Instituto de Investigación en Recursos Cinegéticos (IREC), sede Albacete
Instituto de Investigación en Recursos Cinegéticos (IREC), sede Ciudad Real
Instituto de Investigaciones Marinas (IIM), Vigo
Instituto de Recursos Naturales (IRN), Madrid
Instituto Mediterráneo de Estudios Avanzados (IMEDEA), Mallorca
Instituto Pirenaico de Ecología (IPE-H), Huesca
Instituto Pirenaico de Ecología (IPE-Z), Zaragoza
Museo Nacional de Ciencias Naturales (MNCN), Madrid
Real Jardín Botánico (RJB), Madrid
Unidad de Tecnología Marina (UTM), Barcelona

Área 4. Ciencias Agrarias

Centro de Ciencias Medioambientales (CCMA), Madrid
Centro de Edafología y Biología Aplicada del Segura (CEBAS),Murcia
Estación Agrícola Experimental (EAE), León
Estación Experimental Aula Dei (EEAD), Zaragoza
Estación Experimental del Zaidín (EEZ), Granada
Estación Experimental La Mayora (EELM), Málaga
Instituto de Agricultura Sostenible (IAS), Córdoba
Instituto de Agrobiotecnología (IdAB), Navarra
Instituto de Ciencias Agrarias (ICA), Madrid
Instituto de Ganadería de Montaña (IGM), León
Instituto de Investigaciones Agrobiológicas de Galicia (IIAG), Santiago de Compostela
Instituto de Recursos Naturales y Agrobiología (IRNAS), Sevilla
Instituto de Recursos Naturales y Agrobiología (IRNASA), Salamanca
Misión Biológica de Galicia (MBG), Pontevedra

Área 5. Ciencia y Tecnologías Físicas

Centro de Astrobiología (CAB), Madrid
Centro Nacional de Aceleradores (CNA), Sevilla
Centro Técnico de Informática (CTI), Madrid
Instituto de Acústica (IA), Madrid
Instituto de Astrofísica de Andalucía (IAA), Granada
Instituto de Astronomía y Geodesia (IAG), Madrid
Instituto de Automática Industrial (IAI), Madrid
Instituto de Ciencias del Espacio (ICE), Barcelona
Instituto de Ciencias Matemáticas (ICMAT), Madrid
Instituto de Estructura de la Materia (IEM), Madrid
Instituto de Física Aplicada (IFA) , Madrid
Instituto de Física Corpuscular (IFIC), Valencia
Instituto de Física de Cantabria (IFCA), Santander
Instituto de Física Fundamental (IFF), Madrid
Instituto de Física Interdisciplinar y Sistemas Complejos (IFISC), Palma de Mallorca
Instituto de Física Teórica (IFT), Madrid
Instituto de Investigación en Inteligencia Artificial (IIIA), Barcelona
Instituto de Matemáticas y Física Fundamental (IMAFF), Madrid
Instituto de Microelectrónica (IMM-CNM), Madrid
Instituto de Microelectrónica de Barcelona (IMB-CNM)
Instituto de Microelectrónica de Sevilla (IMSE-CNM)
Instituto de Óptica "Daza Valdés" (IO), Madrid
Instituto de Robótica e Informática Industrial (IRI), Barcelona
Laboratorio de Física de Sistemas Pequeños y Nanotecnología (FSP), Madrid

Laboratorio de Investigación en Tecnologías de la Combustión (LITEC), Zaragoza
Observatorio de Física Cósmica del Ebro (OE), Tarragona
Área 6. Ciencia y Tecnología de Materiales
Centro de Física de Materiales (CFM), San Sebastián
Centro de Investigación de Nanomateriales y Nanotecnología (CINN), Oviedo
Centro de Investigación en Nanociencia y Nanotecnología (CIN2), Barcelona
Centro Nacional de Investigaciones Metalúrgicas (CENIM), Madrid
Instituto de Cerámica y Vidrio (ICV), Madrid
Instituto de Ciencia de Materiales de Aragón (ICMA), Zaragoza
Instituto de Ciencia de Materiales de Barcelona (ICMAB)
Instituto de Ciencia de Materiales de Madrid (ICMM)
Instituto de Ciencia de Materiales de Sevilla (ICMS)
Instituto de Ciencia y Tecnología de Polímeros (ICTP), Madrid
Instituto de Ciencias de la Construcción "Eduardo Torroja" (IETcc), Madrid
Área 7. Ciencia y Tecnología de Alimentos
Instituto de Agroquímica y Tecnología de los Alimentos (IATA), Valencia
Instituto de la Grasa (IG), Sevilla
Instituto de Ciencia y Tecnología de Alimentos y Nutrición (ICTAN), Madrid
Instituto de Fermentaciones Industriales (IFI), Madrid
Instituto de las Ciencias de la Vid y el Vino (ICVV), Logroño
Instituto de Nutrición y Bromatología (INB), Madrid
Instituto de Productos Lácteos de Asturias (IPLA), Villaviciosa
Instituto del Frío (IF), Madrid
Área 8. Ciencia y Tecnologías Químicas
Centro de Investigación y Desarrollo (CID), Barcelona
Centro de investigaciones Científicas Isla de La Cartuja (CARTUJA), Sevilla
Centro de Química Orgánica "Manuel Lora Tamayo" (CENQUIOR), Madrid
Instituto de Carboquímica (ICB), Zaragoza
Instituto de Catálisis y Petroleoquímica (ICP), Madrid
Instituto de Diagnóstico Ambiental y Estudios del Agua (IDAEA), Barcelona
Instituto de Investigaciones Químicas (IIQ), Sevilla
Instituto de Investigaciones Químicas y Ambientales (IIQABD), Barcelona
Instituto de Productos Naturales y Agrobiología (IPNA), Santa Cruz de Tenerife
Instituto de Química Avanzada de Cataluña (IQAC) Barcelona
Instituto de Química Física "Rocasolano" (IQFR), Madrid
Instituto de Química Médica (IQM), Madrid
Instituto de Química Orgánica General (IQOG), Madrid
Instituto de Tecnología Química (ITQ), Valencia
Instituto Nacional del Carbón (INCAR), Oviedo

7. Unravelling the performance of individual scholars: use of Canonical Biplot analysis to explore the performance of scientists by academic rank and scientific field

Artículo aceptado para su publicación en Journal of Informetrics. Autores: Adrián A. Díaz-Faes, Rodrigo Costas, Mª Purificación Galindo, & María Bordons.

Abstract: Individual research performance needs to be addressed by means of a diverse set of indicators capturing the multidimensional framework of science. In this context, Biplot methods emerge as powerful and reliable visualization tools similar to a scatterplot but capturing the multivariate covariance structures among bibliometric indicators. In this paper, we introduce the Canonical Biplot technique to explore differences in the scientific performance of Spanish CSIC researchers, organised by field (Chemistry and Materials Science) and grouped by academic rank (research fellows and three types of full-time permanent scientists). This method enables us to build a Biplot where the groups of individuals are sorted out by the maximum discriminating power between the different indicators considered. Besides, as confidence intervals are displayed in the plot, statistical differences between groups are liable to be studied simultaneously. Since test hypotheses are sensitive to different sample size effects, sizes for some pairwise comparisons are computed. We have found two gradients: a primary gradient where scientists mainly differ in terms of age, production, number of collaborators, number of highly-cited papers and their position in the byline of the publications; and a second gradient, in which scientists with the same academic rank differ by sort of field.

Keywords: Canonical Biplot, multivariate analysis, bibliometrics, individual-level, academic rank.

7.1. Introduction

Scientists form the core of any research system. For this reason, increasing our knowledge on the behaviour and performance of scientists as well as on how they may be influenced by personal characteristics (Costas, van Leeuwen, & Bordons, 2010; Bozeman & Gaughan, 2011) has become an issue of great importance. Such knowledge would be helpful to better understand the research process, inform policy, and improve the ways in which scientists are considered and evaluated in their respective countries and organisations. From a bibliometric standpoint, indicators based on publications can effectively support the assessment process of individual-scholar

performance, although their use at this level is not free of limitations and problems (Radicchi & Castellano, 2013). Specifically, special caution is required in the collection of data and the calculation of indicators due to the difficulties concerning the correct identification of the entire production of scientists and the lower validity of statistical analyses applied to small units. On the other hand, since science has become more complex over the years and demands scientists with different skills and specialisations, there is a large number of factors which should be borne in mind when analysing individual scholars. As a result, the number of bibliometric indicators has significantly grown in recent times in an effort to capture the multidimensionality of scientific activity. Even though since the appearance of the h-index (Hirsch, 2005) there have also been some attempts to shrink different aspects of the research performance into just one bibliometric indicator, the prevalent current belief is that an assorted set of indicators is essential in order to capture the multidimensional nature of academic activity (Moravsick, 1984; Martin, 1996; Costas et al., 2010; Seiler & Wohlrabe, 2013).

Regarding the research performance of scientists, different dimensions such as collaboration, productivity or impact are analysed in bibliometric studies. Thus, inter-field differences in productivity and impact measures and how they may be affected by the basic or applied nature of research have attracted considerable attention in the literature (Bales et al., 2014). Moreover, since teamwork has become a distinctive feature of modern science, the benefits of scientific collaboration in impact and productivity terms has been addressed in several studies (Abramo, D'Angelo, & Solazzi, 2011), as well as the benefits of heterogeneous (Franceschet & Constantini, 2010) and interdisciplinary collaboration, where different participants apply their own insights in a collaborative framework giving way to new knowledge (Sonnenwald, 2007). Concerning collaboration, another factor which has received special consideration has been the position of authors in the byline of scientific publications since it is determined in many disciplines according to the role played by each author in the research (Tscharntke, Hochberg, Rand, Resh, & Kruass, 2007; Waltman, 2012; Liu & Fang, 2014). First and last positions are broadly considered among experimental sciences as the most important places. First-author positions are usually occupied by those in charge of experimental work while last-author positions are reserved for those responsible for the supervision of the research. Finally, personal, institutional and environmental factors may influence the performance of research and have been examined in different studies. In particular, the role played by personal features, such as age or gender, on productivity and research impact and their relation with career success has been explored elsewhere in the literature (Costas et al., 2010; Bozeman & Gaughan, 2011; Costas & Bordons, 2011, Abramo, Cicero, & D'Angelo, 2014).

The behaviour of scientists and its relationship with personal factors may vary between fields. Exploring inter-field differences is a matter of great concern, although it still remains a challenge for bibliometrics due to field-specific productivity and

citation practices (Abramo, Cicero, & D'Angelo, 2013). Interestingly, a recent paper by Ruiz-Castillo and Costas (2014) comprising a vast set of authors shows that despite great differences in productivity, it seems that their distributions are similar across fields suggesting that a single explanation of in-field variation of scientists productivity may suffice. However, from a more theoretical perspective, Bonaccorsi (2008) thoroughly proves that the nature of scientific fields and the problems they face are dissimilar between old and new science fields. In this sense, old science fields like Chemistry deal with simple phenomena of increasing difficulty or with complex phenomena whose fundamental laws are well-known. On the other hand, new fields like Materials Science are placed in a region with high rates of both uncertainty and complexity. Old and new fields may differ in growth rates, degree of diversity, and types of complementarity (resources required as inputs) and this may affect the research performance of scientists. Accordingly, the mainstreaming of these new/old typologies in the analysis of individual scientific performance may also prove relevant.

On the basis of the foregoing, since scientists operate in a complex and multidimensional environment where many factors bear on their scientific activity, a large set of variables is essential for outlining their research performance. In this context, multivariate analysis is useful to provide a reliable picture of the different aspects at play in the activity of individual scholars. In many cases, data structure and relations are far too complex to be successfully addressed through univariate or bivariate methods. For this reason, when we move from a one-dimensional space into a multidimensional one some clear prima facie relations become blurred, but the fact is that one-dimensional approaches are likely to provide inaccurate results since patterns are complex and cannot be described unambiguously (Moravcsik, 1984).

In this paper, we consider that given a set of scientists, grouped by any specific characteristic (e.g. age, academic rank, gender), and a large set of bibliometric indicators that describe their activity, it would be relevant to be able to find out which set of bibliometric indicators discriminate best among the different groups. For this purpose we have chosen Biplot methods since they are powerful and rich visualisation tools, similar to a scatterplot but capturing the multivariate covariance structures among bibliometric indicators. Biplot representations were originally proposed by Gabriel (1971) as methods of multivariate analysis which provide a joint plot of rows and columns in a low-dimensional Euclidean space using markers (points/vectors) for each of them chosen in such a way that the scalar product represents the elements of the data matrix. The markers are obtained by the usual *singular value decomposition* (SVD) of a matrix X_{nxp}, where *n* usually refers to the number of rows (elements) and *p* represents the number of columns (variables) measured on them. Then, the matrix is factorised in row and column markers. Since there are multiple ways to factorise a data matrix, the Biplot representation will have different properties according to the metric selected. They can disclose patterns of covariation and correlation, differences

between groups of sample units and, most importantly, the relation between individual units and the multivariate structure of the data (Gabriel & Odoroff, 1990). Gabriel basically described two types of Biplots (classical Biplots): JK-Biplot in which only the rows are represented with high quality (row-metric preserving), and GH-Biplot in which only the columns are represented with high quality (column-metric preserving). A comprehensive study of the different focuses and alternative types of the Biplot can be found in Cárdenas, Galindo and Vicente-Villardón (2007). In as far as the field of bibliometrics is concerned, Biplot methods were introduced by Díaz-Faes et al. (2011, 2013) and their practical usefulness has been proved in a previous work which explore the performance of networking research centres (Morillo, Díaz-Faes, González-Albo, & Moreno, 2014). These papers consider the HJ-Biplot (Galindo, 1986), which takes the good properties of JK-Biplot and GH-Biplot and provides an optimum quality of representation (QLR) for both points and vectors in the same Cartesian system. Although focused on the JK-Biplot method, another recent study has also discussed its potential in the field of bibliometrics (Torres-Salinas, Robinson-Garcia, Jiménez-Contreras, Herrera, & Delgado López-Cózar, 2013).

In this work, we aim to introduce a new version called 'Canonical Biplot' (Gabriel, 1972; Vicente-Villardón, 1992), which maintain the intrinsic features of the Biplot analysis and allows for a statistical exploration to find out whether there are significant differences in mean values between groups of authors and to identify which bibliometric indicators account for such differences. An appealing application in geology can be found in Varas, Vicente-Tavera, Molina and Vicente-Villardón (2005). Accordingly, this version provides a simultaneous representation of rows (groups) and columns (indicators) in such a way that the groupings are separated by the largest discriminating power between them. In this study, we explore individual performance from different perspectives: collaboration, research level (basic/applied), interdisciplinarity, impact, production, authorship position, academic rank, age and discipline. Finally, some research questions are addressed by means of the integrated analysis of groups and variables: which variables discriminate best between groups of individuals and research fields? Can we anticipate specific patterns or any particular role for scientists based on their academic rank? Which bibliometric indicators best define each group?

7.2. Methods

7.2.1. Data and bibliometric indicators

This study is based on the analysis of the scientific publications of 729 active scientists affiliated to the Spanish National Research Council (CSIC) in 2007 in the research

areas[19] of 'Chemistry Science and Technology' [hereinafter referred to as 'Chemistry'] and 'Materials Science and Technology' [hereinafter referred to as 'Materials Science']. All active scientists included in this study are grouped according to their academic rank in 2007 as defined by the CSIC: 'Post-doc' researchers (which includes various types of research fellows), 'Tenured Scientists' (lowest tenured rank), 'Research Scientists' (intermediate tenured rank), and 'Research Professors' (highest tenured rank). Personal data such as full name, age and academic rank were provided by the CSIC. Table 7.1 shows a summary of the number of individuals in each different group by rank and research area.

Table 7.1. Number of individuals by area and rank

	Chemistry	**Materials Science**	**Total**
Post-doc	26	36	62
Tenured scientist	160	179	339
Research scientist	85	82	167
Research professor	72	89	161
Total	343	386	729

Publications from the Web of Science (WoS) database during the period 2007-2011 were collected for all the individuals included in this study. For a proper allocation of scientific production to individual scholars, an ad hoc software designed to cope with inconsistencies regarding author's names was used in order to automatically collect the publications of the researchers included in this study. A manual revision was conducted to correct minor inconsistencies. As regards citations, this analysis focuses on articles and reviews (hereinafter referred to as "papers") and citations to these papers were collected up to 2013.

Based on all the information above, the following set of indicators was obtained for each researcher:

- ✓ *Age*: age of the researcher at the beginning of period (i.e. 2007).
- ✓ *Production:* the total number of papers published by a researcher during the entire period (2007-2011).

[19] The CSIC comprises approximately 130 research institutes grouped into eight different research areas for the sake of research management (http://www.csic.es/web/guest/areas-cientificas). Every CSIC scholar is assigned to one of these areas.

- ✓ *Pratt index*: measures the concentration of any given researcher papers by subject categories according to the assignment of the publication journals to WoS categories (Pratt, 1977).

$$C = \frac{2\left(\frac{s+1}{2} - q\right)}{s-1}$$

with

$$q = \sum i a_i / t$$

- ✓ Where *s* represents the number of subject categories of the publications, a_i is the size of the category of rank i and t is the total number of publications. It ranges from *C* = 0 which corresponds to a uniform distribution of papers across the categories, to *C* = 1 that corresponds to the case where all papers belong to exactly one category. The lower the concentration, the higher is the interdisciplinarity of the researcher. It is a small variant (normalisation) of the better known Gini index (Carpenter, 1979).

- ✓ *Research level (RL)*: a classification scheme of four research levels was used to describe the basic or applied orientation of a journal. Journals were assigned to a research level on the basis of both expert review and patterns of journal-to-journal citation. It ranges from 1 to 4, where 1 represents the highest level of applied research and 4 refers to the most basic level of research. This classification was described by CHI Research/Computer Horizons Inc. (Narin, Pinski, & Gee, 1976), which now operates as ipIQ. A mean value was assigned to each researcher according to his/her publication pattern.

- ✓ *G-index*: given a set of papers ranked in decreasing order by the number of citations, *g* is the highest (unique) number of papers that together received g^2 or more citations (Egghe, 2006). It combines a measure of quantity (articles) and impact (citations). It was proposed as an improvement of the h-index, since it discriminates better for highly skewed citation distributions.

- ✓ *First author*: proportion of papers that each researcher has published as first author.

- ✓ *Last author*: proportion of papers that each researcher has published as last author[20].

- ✓ *Number of collaborators*: this indicator reflects the distinct number of co-authors with whom one author has published.

[20] Single-authored papers are included in both, first and last author indicators, but only account for 1% of the total number of publications.

- ✓ *Co-authorship index*: this indicator quantifies the average number of authors per publication of a given researcher.

- ✓ *Ptop10%*: this indicator measures the number of papers of a researcher which, compared with other similar papers (same field, publication year, and document type), belong to the top 10% most frequently cited, i.e. Percentile 90.

- ✓ *Mean Normalised Citation Score (MNCS)*: measures the average normalised number of citations received by the papers of a researcher.

$$MNCS = \frac{1}{n}\sum_{i=1}^{n}\frac{c_i}{e_i}$$

  Where *n* is the number of papers of a researcher, c_i denotes the number of citations received by paper *i* and e_i represents the average number of citations of all WoS publications published in the same subject category, in the same year and that have the same document type as paper *i* (Waltman, van Eck, van Leeuwen, Visser, & van Raan, 2011). An MNCS = 1 can be interpreted as the world average, so if MNCS > 1 for a given researcher, it means that he/she has received on average more citations than the world average and vice versa.

- ✓ *Mean Normalised Journal Impact (MNJS)*: is the average normalised citation score of the journals in which a CSIC researcher has published compared to the world average in the same field, year, and document type. Like MNCS, the MNJS indicators corrects for difference among fields.

- ✓ *Proportion of national collaborative publications (pp nat collab)*: this indicator measures the proportion of papers that a researcher has carried out in collaboration between two or more Spanish organizations. It may also include international collaborative papers.

- ✓ *Proportion of international collaborative publications (pp int collab)*: the proportion of co-authored publications with authors affiliated to two or more countries.

- ✓ *Proportion of collaboration with industry (pp industry)*: this indicator quantifies the proportion of papers that a researcher has co-authored with one or more authors affiliated to an industrial partner (Tijssen, 2011).

- ✓ *Mean Geographical Collaboration Distance (MGCD)*: the geographical collaboration distance is defined as the largest geographical distance between two addresses reported in the publication's address list. If a paper contains just one address, then MGCD = 0. A detailed description of the geocoding procedure can be found in Waltman, Tijssen, & van Eck (2011).

- ✓ *Proportion of long distance collaborative publications (pp long dist collab)*: proportion of author publications that have a geographical collaboration distance ≥ 1,000 km.

7.2.2. Canonical Biplot

Gabriel (1972) and Vicente-Villardón (1992) have proposed the Canonical Biplot as an alternative Biplot technique to other similar techniques such as MANOVA, Discriminant Analysis or Canonical Analysis. Given a data matrix X_{nxp}, suppose that *n* rows (in our case 729 authors) can be divided into *K* clearly differentiated groups (e.g. eight groups in our case, this is the combination of the authors' field and rank categories) with n_k authors in each and we have measured *p* variables (in our case, the above mentioned 17 indicators) for each of them. Taking the matrix of means and the covariance matrices between and within the groups it is possible to build a Biplot where the groups are separated by the highest discriminating power between them. If we project the groups on the plot, the coordinates on the first axis represent the linear combination of variables that produces the largest univariate Snedecor *F* in the ANOVA. Accordingly, the bibliometric indicators with larger *F* values will show the highest differences between the groups. This provides a Biplot representation with the following properties: (1) discriminant coordinates (projections in the directions of largest separation), (2) the Euclidean distance between two means markers approximates the Mahalanobis distance between groups (magnitude of effects) and (3) it enables us to place region predictions on the factorial plot.

To illustrate the interpretation of the Canonical Biplot analysis, we provide an example (Figure 7.1). Row markers (stars) depict the average values of the different groups, while vectors represent the different variables (indicated also under V labels). If a group marker projection is close to a variable (end or prolongation of the vector), it means that the average of group is high for that indicator. Conversely, if it is far away it will only take low values. For instance, in Figure 7.1 group G7 has large mean values for variables V2, V4 and V5 (as it is close to them in the projection), while the same group has low mean values for V3, for which groups G1 and G8 obtain large average values.

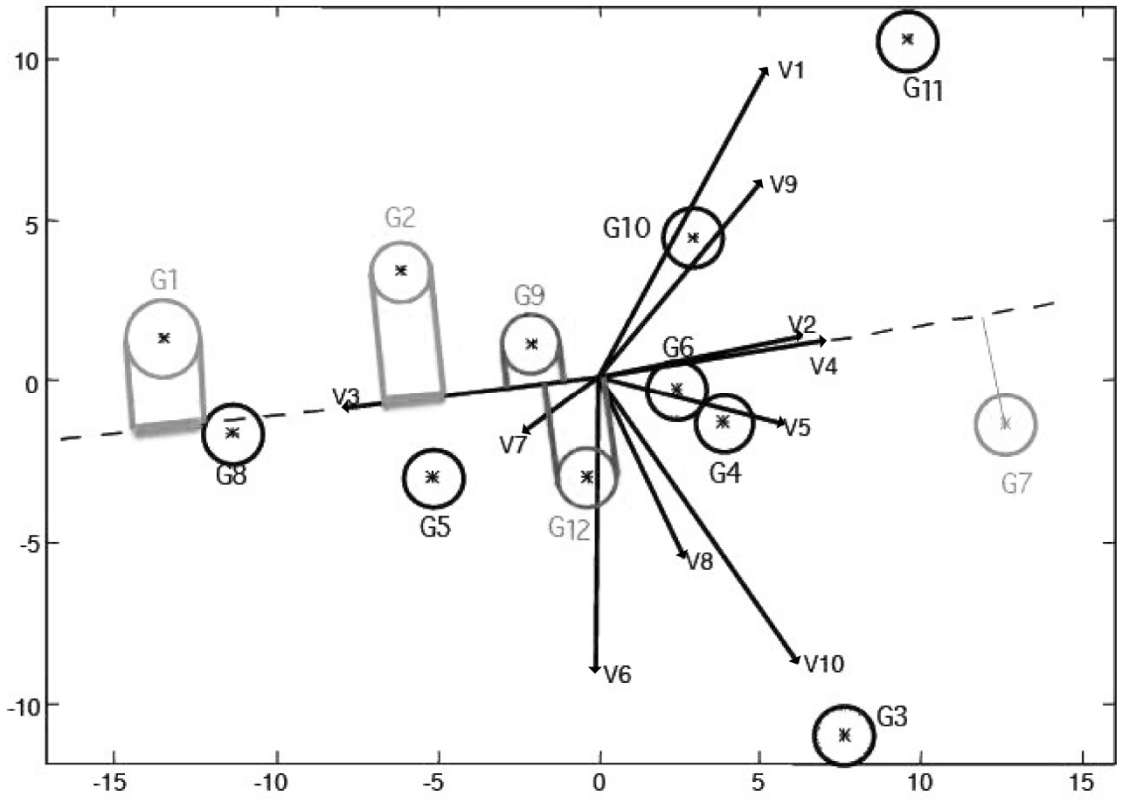


Figure 7.1. Example for the interpretation of the first factorial plane of a Canonical Biplot.

Besides, the angles between vectors (variables) can be interpreted as an approximation of their correlation. Thus, acute angles between two indicators (for instance, V8 and V10, or V2 and V4) indicate that such variables have a strong positive correlation, while obtuse angles (V3 and V4) represent high negative correlations, and right angles (V6 and V4) imply uncorrelated variables. As in other Biplots, we have some measures on the quality of representation and the goodness of fit for bibliometric indicators and, regarding this particular method, for the group means. As the factors obtained are uncorrelated, the inertia (variance) of a variable explained on a particular plane is the sum of the contributions to the axes that form that plane, and that quantity is known as quality of representation (QLR). It is important to bear in mind that only variables and groups with a good QLR should be interpreted on a given plane. In this study, we measure QLR on a 100-point scale.

Confidence intervals are shown along with group's markers in such a way that the projections of the circles onto the direction of a variable approximate a confidence interval. If we project two circles perpendicularly and both intervals do not overlap at all, this indicates that there are significant differences between the means of the groups ($p < 0.05$) (e.g. G1 and G2 over V3 in Figure 7.1). However, should an overlap occur, as is the case between G9 and G12 over V3, we may infer that there are not

significant differences for that variable. The radius of the circle is calculated as $t_{n-K,\alpha/\sqrt{n_K}}$, where $t_{n-K,\alpha}$ is the critical point of a Student's t distribution with $n - K$ degrees of freedom, for a significance level of α (see Amaro, Vicente-Villardón, & Galindo, 2004 for a detailed description). As the radius of the circle is based on a Student's t distribution, a Bonferroni correction, $K = r(r-1)/2$ where K is the number of possible comparisons and r is the number of groups, was added to correct for multiple comparisons which increase the size of the circles and reduce error Type I (wrong rejection of null hypothesis)[21]. Indicators were standardised by columns since they do not have the same measuring scale. The statistical analysis was performed using a Matlab program (Vicente-Villardón, 2014).

7.2.3. Effect sizes

Although our study did not intend to accept or reject null hypothesis for each indicator (from a more classical significant test hypothesis perspective) but to find top discriminant indicators for the different groupings, there are some intrinsic limitations to the results of significance tests described in the literature that we have to bear in mind (Kirk, 1996; Schneider, 2013). In this study, Post-doc groups have a small sample size as compared to permanent scientists (i.e. Tenured Scientists, Research Scientists and Research Professors). Therefore, we may argue that significance tests are more sensitive for Post-docs in terms of wrong acceptance of the null hypothesis (Type II error). Thus, we have calculated the size effect using Hedge's *g* (Hedges, 1981), which is similar to Cohen's *d* (Cohen, 1988) but weighing each group's standard deviation by its sample size. This effect quantifies the magnitude of the differences between groups of scientists expressed in standard deviation units. Since $g \cong d$, we have taken the values proposed by Cohen (1988), which are widely accepted: <0.20 trivial, >= 0.20 small effect, >=0.50 medium effect, >= 0.80 large effect and >= 1.30 very large effect[22]. Effect sizes were computed by means of the R package 'compute.es' (Del Re, 2013).

7.3. Results

CSIC scientists published a total of 9,163 papers in the WoS database during the period 2007-2011 with a resulting breakdown by research area of 4,886 in Materials Science (MAT) and 4,277 in Chemistry (CHEM). Table 7.2 displays the means and standard deviations for the 17 variables analysed by academic rank and field. We provide this information to support our subsequent results.

[21] Note that the method is conservative since where two circles do not overlap we may assume that there is a significant difference, but if there is an overlap we may find a significant difference in another direction of the multidimensional space (Varas et al., 2005).

[22] These thresholds are informative and easy to understand but they should not be used uncritically as benchmarks (Cohen, 1988).

A glimpse at Table 7.2 gives us some preliminary hints about the scientific performance of the researchers under study. For instance, g-index is larger for higher academic positions in both areas, ranging from g = 7 or 8 for Post-docs to g = 13 for Research Professors in both research areas. However, excluding Post-docs, scientists in a given rank show very similar g-index in both areas. When the research level is considered, it tends to be more basic on average for Chemistry scientists regardless the rank when compared to Materials Science. Interestingly, collaboration with industry reaches its highest point at the lowest and intermediate tenured positions (Tenured Scientists and Research Scientists) and shows low dispersion for all scientists in both the Chemistry and Materials areas. MNJS shows that Post-docs publish on average in slightly more outstanding journals in terms of citations, although their values are more scattered across the distribution. All in all, the information in Table 7.2 is highly revealing but also very massive, and it is just here where the use of the Canonical Biplot can help to better summarise and discuss the differences and main patterns among the different indicators and groups of researchers, as described in the section below.

7.3.1. Biplot analysis

The factorial plot resulting from the Canonical Biplot analysis is displayed in Figure 7.2. We analyse the main factorial plane which accounts for most of the variance. The value of Wilk's lambda distribution (λ = 7.914 with $p < 0.01$) shows that the difference between the means of groups actually exists, i.e. there are indeed differences that can be attributed to the groups under survey. The inertia absorption for the first factorial plane is 91.35%. Axis 1 accounts for 80.23% of the variance whereas Axis 2 explains 11.12%.

Table 7.2. Average values for individual-level indicators.

Groups	Age	Pratt Index	Research Level	G-Index	Production	First author	Last author	N Collaborators	Co-authorship index	Ptop10%	MNCS	MNJS	pp int collab	pp nat collab	pp industry	MGCD	pp long dist collab
Post-doc (CHEM)	35 ± 5	.33 ± .20	3.21 ± .68	7 ± 6	10 ± 10	.37 ± .28	.07 ± .11	24 ± 23	5.58 ± 1.27	1.38 ± 4.28	.92 ± .65	1.48 ± .60	.31 ± .25	.56 ± .35	.020 ± .06	918 ± 817	.28 ± .24
Tenure (CHEM)	44 ± 8	.43 ± .18	3.18 ± .75	10± 6	16 ± 12	.15 ± .22	.22 ± .24	36 ± 30	6.02 ± 1.61	2.09 ± 2.97	1.15 ± 1.15	1.35 ± .39	.33 ± .25	.51 ± .30	.042 ± .08	1,592 ± 1,595	.29 ± .24
Researcher (CHEM)	51 ± 7	.45 ± 19	3.33 ± .73	11 ± 8	21 ± 25	.09 ± .17	.36 ± .28	40 ± 33	5.80 ± 1.56	2.37 ± 3.95	.99 ± .66	1.36 ± .44	.34 ± .26	.54 ± .32	.041 ± .11	1,662 ± 1,829	.29 ± .25
Professor (CHEM)	59 ± 7	.47 ± .20	3.32 ± .59	13 ± 12	30 ± 46	.08 ± .17	.32 ± .26	60 ± 79	5.60 ± 1.31	5.62 ± 15.44	1.01 ± .73	1.32 ± .42	.33 ± .24	.53 ± .31	.030 ± .05	1,640 ± 1,600	.29 ± .23
Post-doc (MAT)	35 ± 3	.40 ± .16	3.03 ± .65	8 ± 4	11 ± 6	.31 ± .22	.06 ± .09	24 ± 15	5.81 ± 1.77	.90 ± .94	1.05 ± .52	1.41 ± .58	.39 ± .29	.49 ± .34	.024 ± .05	1,189 ± 1,169	.31 ± .25
Tenure (MAT)	45 ± 8	.46 ± .19	2.95 ± .61	10 ± 6	18 ± 11	.16 ± .15	.18 ± .15	39 ± 28	5.46 ± 1.50	1.95 ± 2.35	1.04 ± .79	1.26 ± .54	.42 ± 27	.49 ± .30	.037 ± .09	1,869 ± 1,595	.36 ± .24
Researcher (MAT)	51 ± 8	.46 ± .17	3.00 ± .62	11 ± 6	21 ± 13	.14 ± .16	.24 ± .19	43 ± 31	5.31 ± 1.27	2.31 ± 3.31	1.00 ± .76	1.26 ± .49	.39 ± .22	.54 ± .28	.034 ± .06	1,632 ± 1,482	.32 ± .20
Professor (MAT)	57 ± 7	.50 ± .18	3.14 ± .58	13 ± 10	30 ± 23	.08 ± .09	.30 ± .17	61 ± 52	5.58 ± 1.38	3.46 ± 6.40	1.14 ± 1.52	1.32 ± .54	.45 ± .25	.51 ± .27	.024 ± .04	2,003 ± 1,584	.37 ± .23
Total CHEM	48 ± 10	.44 ± .19	3.3 ± .71	11 ± 8	20 ± 27	.14 ± .22	.27± .26	41 ± 46	5.84 ± 1.52	2.85 ± 7.8	1.07 ± .77	1.36 ± .43	33 ± .25	.53 ± .31	.037 ± .08	1,568 ± 1619	29 ± .24
Total MAT	48 ± 10	.44 ± .19	3 ± .61	11 ± 7	21 ± 16	.15 ± .16	.21 ± .17	44 ± 36	5.50 ± 1.46	2.85 ± 7.8	1.06 ± .98	1.29 ± .53	.42 ± .26	.51 ± .29	.032 ± .07	1,786 ± 1546	35 ± .23

Note: Data expressed as average ± standard deviation.

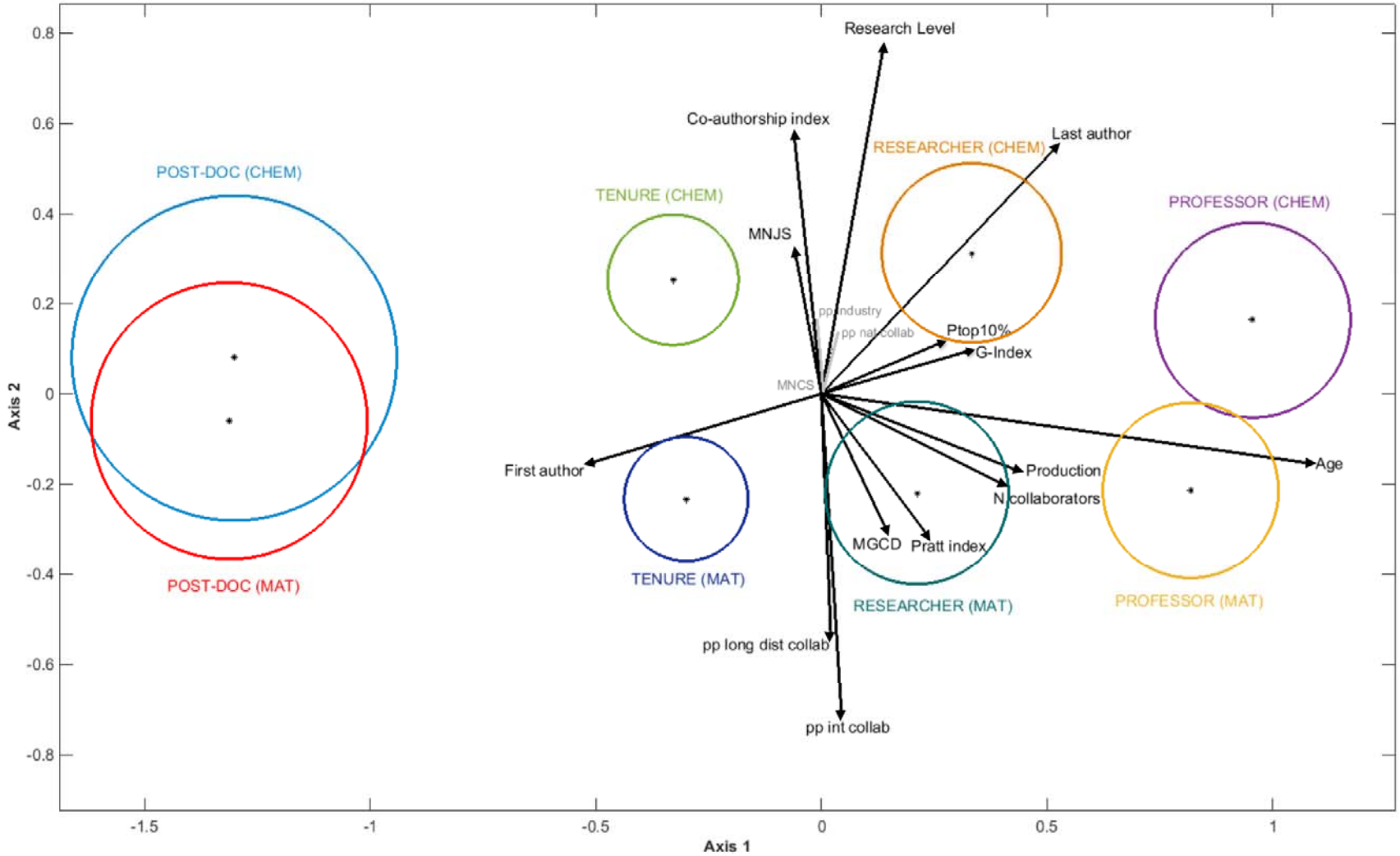


Figure 7.2. Two dimensional Canonical Biplot for authors performance (plane 1-2). Variables coloured in grey present a QLR below 50 points.

Since Canonical Biplots provide discriminant coordinates, a straightforward separation among the groupings of scientists can be clearly seen in the first factorial plane. The most remarkable feature is that there are major discrepancies between scientists from different academic ranks since they show singular features which separate them from the rest, especially in the lowest and highest ranks. There are global differences in the ANOVA test ($p < 0.01$) for all indicators except for pp nat collab, pp industry and MNCS. The coordinates of Axis 1 show that the discrepancies between Post-docs and the three types of full-time scientists are mainly based on age, production, number of collaborators and order of signature (first author, last author). These indicators have the largest univariate Snedecor *F* in the ANOVA. Accordingly, the variables that best separate one group from another present high contributions to the axes and those who present low discriminatory power have a poor QLR (coloured in grey in Figure 7.2) and their information is put aside by the Biplot (Table 7.3). For further analysis on the discrepancies between pairs of groups, we focus on the projections of the circles in the direction of the variables.

Table 7.3. Quality of representation (QLR) to explain the group means.

Variables	Axis 1	Axis 2	Axis 3	Axis 4	Axis 5
Age	**99.40**	0.28	0.13	0.13	0.01
pp nat collab	11.48	21.25	17.59	14.17	14.08
Pratt index	**65.55**	16.09	13.73	0.34	2.15
Research level	16.72	**72.53**	7.16	2.13	0.58
G-Index	**96.00**	1.11	0.21	0.15	1.49
Production	**92.91**	1.96	2.38	0.00	1.55
First author	**80.25**	0.99	17.6	0.77	0.07
Last author	**79.35**	12.14	3.00	3.47	1.52
N collaborators	**88.68**	3.03	1.3	1.42	5.42
Co-authorship index	5.54	**69.12**	5.45	0.00	19.25
Ptop10%	**70.89**	1.78	13.12	11.02	0.68
MNCS	0.03	1.41	27.5	4.17	**60.85**
MNJS	10.25	41.91	25.79	11.19	9.82
pp int collab	2.33	**82.43**	4.85	4.9	2.91
pp industry	1.58	19.30	**58.66**	6.75	13.65
MGCD	34.68	20.98	34.44	0.91	2.25
pp long dist collab	0.85	**78.37**	10.32	0.31	3.85

* Bold values > 50.

When the circles of the groups are projected onto different vectors, we observe for Axis 1 that there are statistical differences (the projections of the circles do not overlap) on age, production, g-index, number of collaborators, ptop10% and the proportion of papers published as first author for each academic rank. Thus, as we move from left to right across the plot, researchers with the highest rank (Research

Professors) are older and show higher rates of performance with a growing number of papers published, their networks in terms of collaborators become larger and their rate of papers with a high number of citations (acute angle between vectors) is higher. On the other hand, as scientists move up in academic rank, interdisciplinarity (Pratt index) tends to shrink and their publications are concentrated in a smaller number of subject categories. Interestingly, the measures of individual citation performance vary according to the selected indicator. For size-dependent indicators, such as ptop10% and g-index, a better citations performance is found for higher ranks. Nevertheless, if we take the MNCS for reference, the behaviour on average is similar for all scientists as we have already pointed out (no global differences). Regarding the order of signature, it appears, for instance, that last authors seem to be more common at higher academic ranks, although there are no significant differences for some pairwise comparisons such as between Researcher Scientists and Research Professors in Chemistry (projections of circle do overlap). On the other hand, Post-docs most distinctive feature, which separate them from the rest of the groups, is the higher proportion of papers signed as first authors. This variable has a high negative correlation (obtuse angles) with the aforesaid indicators for which scientists with high academic rank show top values. All in all, the first position of signature seems to discriminate very well between low and high academic ranks in both fields.

Concerning Axis 2, we observe there are inherent features among full permanent scientists for each field due to the fact that scientists within the same rank perform alike in both fields (Figure 7.2). Chemistry scientists are placed on top of the plot performing on average a more basic and slightly more interdisciplinary research with a higher number of collaborators (group markers are close in projection to research level and the co-authorship index and far away from the Pratt index). Materials Science researchers, which are placed at the bottom, collaborate more at international level with more distant partners and are involved in more applied research. Moreover, a higher MNJS is observed in Chemistry when compared to that in Materials Science.

On the other hand, we have also depicted plane 3-5 where most of the variance for collaboration with industry and MNCS is explained (Figure 7.3). This plane can be interesting in order to observe that these two indicators show a weak discriminatory power. Axis 3 is chiefly characterised by pp industry whereas Axis 5 accounts for MNCS. It is clear that all confidence intervals overlap, so there is no difference between groupings in terms of average citations and the degree of collaboration with enterprises. Besides, there is no relation between both indicators (right angle between vectors).

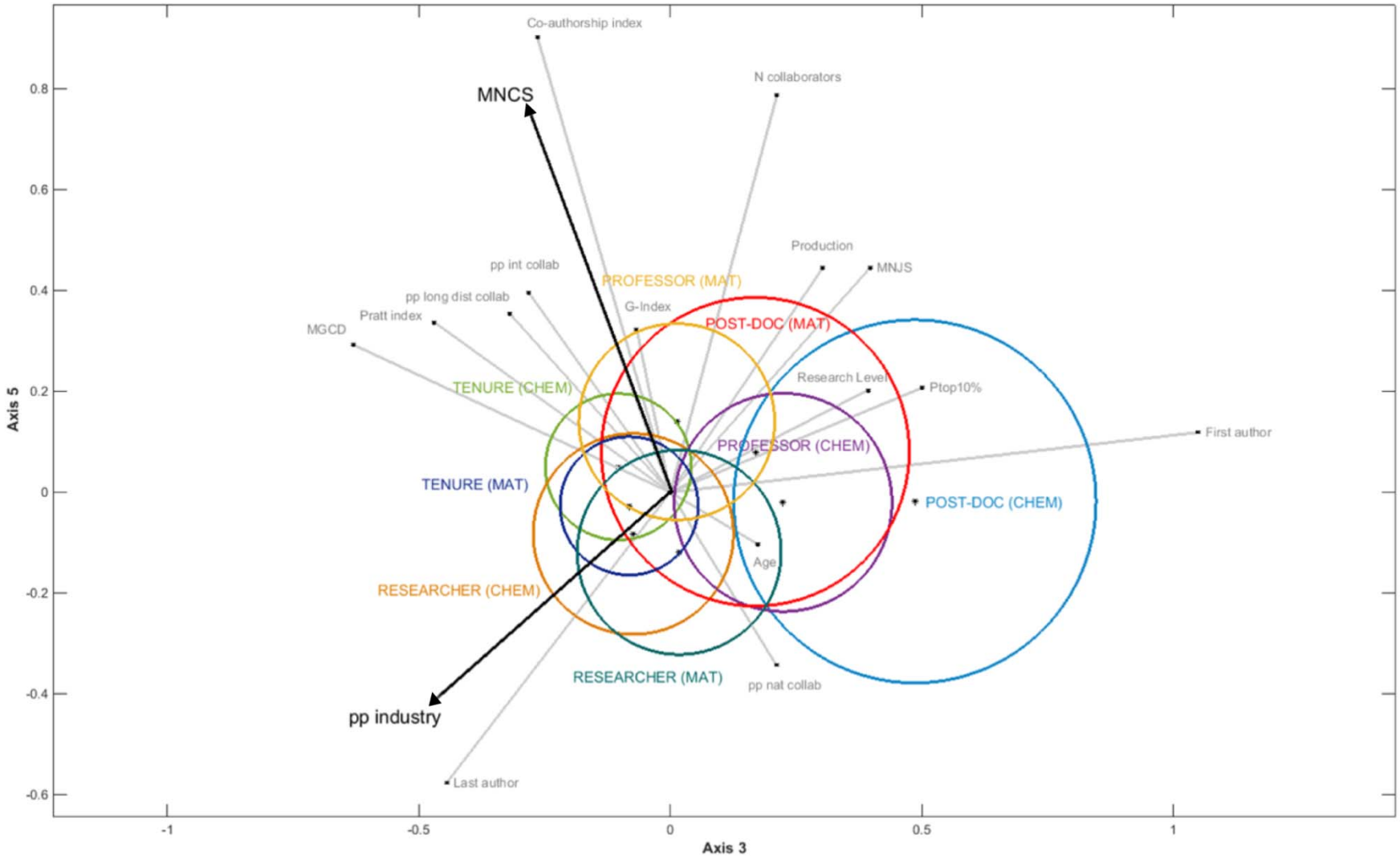


Figure 7.3. Two dimensional Canonical Biplot for authors performance (plane 3-5). Variables coloured in grey present a QLR below 50 points.

7.3.2. Effect sizes results

We have computed effect sizes by means of Hedge's *g* for some pairwise comparisons in order to estimate the magnitudes of the differences between groupings. These results are useful both due to the influence of sample size in the statistical values and as a means to support the Biplot analysis. Since the number of possible comparisons is extremely high, Table 7.4 displays a small but meaningful selection of effect sizes and variables. We mainly display effect sizes for Post-docs because of their smaller sample size.

Table 7.4. Hedge's *g* by field and academic rank.

Academic rank	Production	N Collaborators	First Author	G-index	pp int collab	Research Level	MNCS	pp industry
Post-doc (CHEM) vs. Tenure (CHEM)	0.49	0.43	**0.96**	0.57	0.05	0.04	0.27	0.30
Post-doc (CHEM) vs. Researcher (CHEM)	0.52	0.52	**1.35**	0.60	0.12	0.16	0.11	0.20
Post-doc (CHEM) vs. Professor (CHEM)	0.52	0.53	**1.41**	0.58	0.06	0.18	0.14	0.20
Tenure (CHEM) vs. Professor (MAT)	**0.83**	0.63	0.34	0.32	0.47	0.07	0.01	0.26
Researcher (CHEM) vs. Tenure (MAT)	0.20	0.04	0.40	0.26	0.17	0.58	0.07	0.04
Post-doc (MAT) vs. Tenure (MAT)	0.62	0.58	**0.94**	0.35	0.13	0.13	0.00	0.16
Post-doc (MAT) vs. Researcher (MAT)	**0.81**	0.71	**0.94**	0.56	0.03	0.05	0.07	0.17
Post-doc (MAT) vs. Professor (MAT)	**0.91**	**0.83**	**1.60**	0.57	0.23	0.17	0.07	0.01
Tenure (MAT) vs. Professor (CHEM)	0.47	0.44	0.50	0.43	0.35	0.61	0.05	0.08
Researcher (MAT) vs. Professor (CHEM)	0.29	0.30	0.36	0.23	0.27	0.53	0.01	0.06
Post-doc (MAT) vs. Professor (CHEM)	0.50	0.56	**1.22**	0.53	0.22	0.47	0.06	0.12
Post-doc (CHEM) vs. Professor (MAT)	**0.94**	0.78	**1.83**	0.64	0.53	0.12	0.16	0.10

* Effect size: *g* <0.20 trivial, *g* >= 0.20 small effect, *g* >=0.50 medium effect, *g* >= 0.80 large effect and *g* >= 1.30 very large effect.

Accordingly, the indicators which showed a stronger discriminatory power among groups of authors such as production, number of collaborators, first author and g-index obtain high *g* values as well. We can see what variables and pairs of groups differ most in terms of standard deviation units. In this regard, Post-doc researchers presented as a singular feature appear as first author in the byline at the Canonical Biplot. This becomes clear looking at Hedge's *g* for Post-docs with regard to the three types of full permanent scientists since size effects are larger ($g > 0.80$), which means the distribution of group scores do not overlap. Likewise, production, number of collaborators and g-index range from medium to large effect sizes which can be considered consistent enough. For instance, the largest effect sizes for production are

observed between Post-docs and Research Professors, but also for Tenure (CHEM) vs. Professor (MAT) where g = 0.83. Interestingly, if we focus on international collaboration and research level, which were variables accounted for on Axis 2 of the Biplot, distributions within the same field are similar for all four academic ranks. Nevertheless, the biggest effect sizes arise when scientists from distinct fields are compared (Post-doc (CHEM) vs. Professor (MAT) g = 0.53 for pp int collab and Tenure (MAT) vs. Professor (CHEM) g = 0.61 for research level). On the other hand, MNCS and collaboration with industry, two of the indicators which showed weak discriminatory power for group means (no global differences) also present small g values for all pairwise comparisons, i.e. the differences are scant between groupings.

7.4. Discussion and conclusions

Due to the multidimensional nature of research activity, multiple indicators are needed to obtain a reliable picture of the research performance of scientists. Moreover, research performance may be affected by a wide range of factors, including personal data (age, gender), career (academic rank, promotion), organisational (teaching, mobility) or economic issues (research funds); and this generates a wide diversity of possibilities for the analysis of scientific performance. Due to this diversity and its multidimensionality, it is sometimes difficult to extract meaningful information from a multiplicity of variables and groups of researchers. In this study we provide a practical solution to this problem. Canonical Biplot analysis has been applied to study the behaviour of 729 researchers in two areas at the Spanish CSIC. Our aim was to build a joint plot for 8 groups (2 fields x 4 academic ranks) and 17 variables where the groups were separated by the maximum discriminating power between them. As a result, we have studied which structure of variables offers the highest separation between groups of authors in a broad set of bibliometric indicators. Besides, statistical differences between group means have been assessed. Since test hypotheses are sensitive to different samples sizes, we have supported our results by means of effect sizes. They can be a relevant measure at the individual-level since data collection at this level is a tough job and it is not always possible to achieve a good sample size for the different categories or groupings.

In contrast to other multivariate techniques, Canonical Biplot offers some interesting and advantageous features. For instance, if we would have applied a MANOVA, we should have examined many tables and we would not have obtained a joint representation in a low dimensional space for a visual inspection of the underlying structure of the data matrix. If we would have used a Discriminant Analysis, we would have obtained a low dimensional plot describing the group's structure, but we would not have had direct information about the bibliometric indicators responsible for the separation between groupings and their correlations.

In our study, we have found two gradients for CSIC scientists. First, researchers of different academic ranks are clearly separated according to their age, level of production of papers, distinct number of collaborators, the number of highly-cited papers and their position in the byline. The second gradient relates to intrinsic field features since it separates Chemistry from Materials Science. The scientific activity of Materials Science researchers is less interdisciplinary and shows a higher share of international links compared to Chemistry, where the nature of research is more basic and takes place in larger teams. The lower interdisciplinarity of Materials Science is an unexpected finding, since it is a more recent and more applied field and it is supposed to require more diverse inputs. Actually, interdisciplinarity is often associated with application-oriented research and complex problem solving (van Rijnsoever & Hessels, 2011). A possible explanation is that our interdisciplinarity measure considers the diversity of WoS categories to which the publication journals are assigned, but not the heterogeneous nature of each category. Accordingly, taking into account other measures such as the specialisation of team members could be advisable.

Our results show that the higher the academic rank, the higher the production of the researcher and his/her share of highly-cited papers and the more likely it is to appear as last author in the byline. The higher production of researchers in the highest academic rank has been described for other research areas at the Spanish CSIC (Costas et al., 2010) and elsewhere (Abramo, D'Angelo, & Di Costa, 2011). Moreover, Abramo et al. (2014) denoted a moderate correlation for several fields (mild for Chemistry) between being a top productive scientist and the probability of having produced highly-cited papers. The same study also evidences that the highest probability of yielding highly-cited papers for non-top productive scientists lies with Assistant Professors, which can be considered to be on the same level as our Post-doc scientists. Interestingly, although our results show no differences in MNCS by academic rank and field, Post-docs obtain the highest average MNJS, probably because they are at the beginning of their careers and are very aware of the importance of publishing in prestigious journals to build a solid scientific career and gain a tenured position. On the other hand, an enhanced impact has been identified for those scientists who have heterogeneous collaboration patterns, chiefly at the international level (Franceschet & Constantini, 2010; Abramo et al. 2011; Bordons, Aparicio, & Costas, 2013), even though there are some exceptions, maybe due to hyper-authorship (Cronin, 2001). Our results show that there are differences on international and long distance collaboration, but they seem to be more related to field adscription. However, having an extensive network of collaborators proved valuable in terms of performance.

As regards authorship order and age, it has been shown in the literature that, in experimental sciences, first-authored papers predominate among younger researchers while last position tends to be reserved to the more experienced scientists (Gingras, Larivière, Macaluso, & Robitalle, 2008; Costas & Bordons, 2011). Last-author position

has been pointed out as a major position since it can be assumed to name the driving force behind the research, both intellectually and financially (Tscharntke et al., 2007). This is the role mainly played by authors in the higher academic ranks. Nevertheless, the strongest contribution to the actual work carried out is often expected from the researcher in the first-author position of the byline (Liu & Fang, 2014). This is indeed the most distinctive feature of Post-docs in our study, since they obtain the highest share of first-authored papers (around 1/3 of their production) and the lowest proportion of last-authored ones. Post-docs might be able to lead a research line and can work independently or within an existing team. Since we are dealing with two experimental fields, the junior signing pattern observed for Post-docs (Costas & Bordons, 2011) and their relatively high co-authorship index suggest their integration into more or less consolidated teams, which in fact is being fostered by some programmes to increase the size and competitiveness of teams. Although Post-docs are not so well connected to other people as researchers in tenured positions and yield less number of papers, they obtain an outstanding impact as measured by the MNJS. These results are consistent with previous studies focused on academic rank influence and signature order for CSIC scientists (see Costas & Bordons, 2011), where a relation between publishing in prestigious journals or being highly rewarded with citations and any specific position in the byline was not found.

We have to mention some limitations to the present study. First, the fact that each scientist was considered in the professional rank he/she held in 2007 and some of them may have been promoted during our period of reference. Nonetheless, this is only the case in a very small percentage of scientists. Second, due note should be taken that the small sample size of Post-docs limits the scope of the analysis and the significance of the results concerning this category of researchers. And third and last, our study focuses on CSIC scientists in two fields and we cannot generalise our results across other fields, institutions or countries. Concerning future research, including other factors such as promotion, teaching or mobility issues might be helpful to better understand certain patterns with a bearing on performance. PhD students also remain as an attractive group for further study due to the lack of large-scale analyses on them (Larivière, 2012, is an exception).

In summary, the approach set forth in this paper concerning the study of research performance at the individual level by means of a Canonical Biplot analysis enables us to examine a large set of indicators and to explore the underlying matrix structure. In this specific case, the most distinctive patterns that characterise researchers grouped by field and academic rank have been revealed. In view of all the foregoing, we conclude that the Canonical Biplot analysis is a strong exploratory tool with high potential in order to make headway in the unravelling of the intricate structure of relationships between research performance indicators and the individual characteristics of researchers.

Acknowledgments

This research was supported by the Spanish National Research Council (JAE pre-doctoral grant). This paper was conceived while Adrián A. Díaz-Faes carried out a research stay at the Centre for Science and Technology Studies (CWTS), Leiden University (The Netherlands).

References

Abramo, G., D'Angelo, C.A., & Di Costa, F. (2011). Research productivity: Are higher academic ranks more productive than lower ones? *Scientometrics*, *88,* 915–928.

Abramo, G., D'Angelo, C.A., & Solazzi, M. (2011). Are researchers that collaborate more at the international level top performers? An investigation on the Italian university system. *Journal of Informetrics*, *5*, 204-211. doi:10.1016/j.joi.2010.11.002

Abramo, G., Cicero, T., & D'Angelo, C. A. (2013). Individual research performance: A proposal for comparing apples to oranges. *Journal of Informetrics*, *7*(2), 528–539. doi:10.1016/j.joi.2013.01.013

Abramo, G., Cicero, T., & D'Angelo, C.A. (2014). Are authors of highly cited articles also the most productives ones? *Journal of Informetrics*, *8*, 89-97. doi:10.1016/j.joi.2013.10.011

Amaro I.R., Vicente-Villardón, J.L., & Galindo, M.P. (2004). MANOVA Biplot para arreglos de tratamientos con dos factores basado en modelos lineales generales multivariantes. *Interciencia*, *29*(1), 26-32.

Bales, M.E., Dine, D.C., Merrill, J.A., Johnson, S.B., Bakken, S., & Weng. C. (2014). Associating co-authorship patterns with publications in high-impact journals. *Journal of Biomedical Informatics*, *52,* 311-318.

Bonacorsi, A. (2008). Search regimes and the industrial dynamics of science. *Minerva*, *46*, 285-315.

Bordons, M., Aparicio, J., & Costas, R. (2013). Heterogeneity of collaboration and its relationship with research impact in a biomedical field. *Scientometrics*, *96*, 443-466.

Bozeman, B., & Gaughan, M. (2011). How do men and women differ in research collaborations? An analysis of the collaborative motives and strategies of academic researchers. *Research Policy*, *40*, 1393,1402. doi: 10.1016/j.respol.2011.07.002

Cárdenas, O., Galindo, M.P., & Vicente-Villardón, J.L. (2007). Los métodos Biplot: evolución y aplicaciones. *Revista Venezolana de Análisis de Coyuntura*, *13*(1), 279-303.

Carpenter, M.P. (1979). Similarity of Pratt's measure of class concentration to the Gini index. *Journal of the American Society for Information Science*, *30*(2), 108-110.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2ª ed.). Hillsdale, N.J.: Lawrence Erlbaum Associates.

Costas, R., van Leeuwen, T. N., & Bordons, M. (2010). A Bibliometric Classificatory Approach for the Study and Assessment of Research Performance at the Individual Level : The Effects of Age on Productivity and Impact. *Journal of the American Society for Information Science and Technology*, *61*(8), 1564–1581. doi:10.1002/asi.21348

Costas, R., & Bordons, M. (2011). Do age and professional rank influence the order of authorship in scientific publications? Some evidence from a micro-level perspective. *Scientometrics*, *88*, 145-161.

Cronin, B. (2001). Hyperauthorship: A postmodern perversion or evidence of a structural shift in scholarly communication practices? *Journal of the American Society for Information Science and Technology*, *52*(7), 558–569.

Díaz-Faes, A.A., Benito-García, N., Martín-Rodero, H., & Vicente-Villardón, J.L. (2011). Propuesta de aplicabilidad del método multivariante gráfico Biplot a los estudios bibliométricos en biomedicina. *Actas XIV Jornadas Bibliosalud*, p. 66. Cádiz, España: BV-SSPA. Retrieved from: http://hdl.handle.net/10760/15998

Díaz-Faes, A.A., González-Albo, B., Galindo, M. P., & Bordons, M. (2013). HJ-Biplot como herramienta de inspección de matrices de datos bibliométricos. *Revista Española de Documentación Científica*, *36*(1), e001. doi:10.3989/redc.2013.1.988.

Del Re, A.C. (2013). compute.es: Compute Effect Sizes. R package version 0.2-2. Retrieved from: http://cran.r-project.org/web/packages/compute.es

Egghe, L. (2006). Theory and practise of the g-index. *Scientometrics*, *69*(1), 131–152.

Franceschet, M., & Costantini, A. (2010). The effect of scholar collaboration on impact and quality of academic papers. *Journal of Informetrics*, *4*, 540-553.

Gabriel, K.R. (1971). The Biplot graphic display of matrices with application to principal component analysis. *Biometrika*, *58*(3), 453-467.

Gabriel, K.R. (1972). Analysis of meteorological data by means of canonical decomposition and biplots. *Journal of Applied Meteorology*, *11*, 1071–1077.

Gabriel, K.R., & Odoroff, C.L. (1990). Biplots in biomedical research. *Statistics in Medicine*, *9*, 469-485.

Galindo, M.P. (1986). Una alternativa de representación simultánea: HJ-Biplot. *Qüestioó*, *10*(1), 13-23.

Gingras, Y., Lariviere, V., Macaluso, B., & Robitaille, J. P. (2008). The effects of aging on researchers'publication and citation patterns. *PloS One*, *3*(12), e4048.

Hedges, L.V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics*, *6*(2), 106-128.

Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(46), 16569–16572.

Kirk, R.E. (1996). Practical significance: a concept whose time has come. *Educational and Psychological Measurement, 56*(5), 746-759.

Larivière, V. (2012). On the shoulders of students? The contribution of PhD students to the advancement of knowledge. *Scientometrics*, *90*(2), 463–481. doi: 10.1007/s11192-011-0495-6

Liu, X.Z., & Fang, H. (2014). Scientific group leaders' authorship preferences: an empirical investigation. *Scientometrics*, *98*, 909-925.

Martin, B.R. (1996). The use of multiple indicators in the assessment of basic research. *Scientometrics 36*(3), 343-362.

Moravcsik, M.J. (1984). Life in a multidimensional world. *Scientometrics*, *6*(2), 75-86.

Morillo, F., Diaz-Faes, A.A., González-Albo, B., & Moreno, L. (2014). Do networking centres perform better? An exploratory analysis in Psychiatry and Gastroenterology/Hepatology in Spain. *Scientometrics*, *98*(2), 1401-1416. doi: 10.1007/s11192-013-1183-5

Narin, F., Pinski, G., & Gee, H.H. (1976). Structure of the biomedical literature. *JASIS, 27*(1), 25-44.

Pratt, A.D. (1977). A measure of class concentration in bibliometrics. *JASIS*, 28(5), 285-292.

Radicchi, F., & Castellano, C. (2013). Analysis of bibliometric indicators for individual scholars in a large data set. *Scientometrics*, *97*(3), 627–637. doi:10.1007/s11192-013-1027-3

Ruiz-Castillo, J., & Costas, R. (2014). The skewness of scientific productivity. *Journal of Informetrics*, *8*, 917-934.

Schneider, J. W. (2013). Caveats for using statistical significance tests in research assessments. *Journal of Informetrics*, *7*(1), 50–62. doi:10.1016/j.joi.2012.08.005

Seiler, C., & Wohlrabe, K. (2013). Archetypal scientists. *Journal of Informetrics*, *7*(2), 345–356. doi:10.1016/j.joi.2012.11.013

Sonnenwald, D.H. (2007). Scientific collaboration. *Annual Review of Information Science and Technology*, *41*, 643-681.

Tijssen, R.J.W. (2011). Joint research publications: a performance indicator of university-industry collaboration. *Evaluation in Higher Education*, *5*(2), 19-40.

Torres-Salinas, D., Robinson-Garcia, N., Jiménez-Contreras, E., Herrera, F., & Delgado López-Cózar, E. (2013). On the use of biplot analysis for multivariate bibliometric and scientific indicators. *Journal of the American Society for Information Science and Technology. 64*(7), 1468-1479.

Tscharntke, T., Hochberg, M.E., Rand, T.A., Resh, V.H., & Kruass, J. (2007). Author sequence and credit for contributions in multiauthored publications. *Plos Biology*, *5*(1), 13-14.

van Rijnsoever, F.J., & Hessels, L.K. (2011). Factors associated with disciplinary and interdisciplinary research collaboration. *Research Policy*, *40*, 463–472.

Varas, M.L., Vicente-Tavera, S., Molina, E., & Vicente-Villardón (2005). Role of Canonical Biplot method in the study of building stones: an example from Spanish monumental heritage. *Environmetrics*, *16*, 1-15.

Vicente-Villardón J.L. (1992). *Una alternativa a los métodos factoriales clásicos basada en una generalización de los métodos biplot*. PhD thesis. Salamanca, Spain: University of Salamanca.

Vicente-Villardón, J.L. (2014). MultBiplot: A package for Multivariate Analysis using Biplots. Departamento de Estadística, Universidad de Salamanca [Software]. Retrieved from: http://biplot.usal.es/ClassicalBiplot/index.html.

Waltman, L., Tijssen, R.J.W., & van Eck, N.J. (2011). Globalisation of science in kilometres. *Journal of Informetrics*, *5*(4), 574-582. doi:10.1016/j.joi.2011.05.003

Waltman, L., van Eck, N.J., van Leeuwen, T.N., Visser, M.S., & van Raan, A.F.J. (2011). Towards a new crown indicator: Some theoretical considerations. *Journal of Informetrics*, *5*(1), 37-47. doi:10.1007/s11192-011-0354-5

Waltman, L. (2012). An empirical analysis of the use of alphabetical authorship in scientific publishing. *Journal of Informetrics*, *6*(4), 700–711. doi:10.1016/j.joi.2012.07.008

8. The relationship between the research performance of scientists and their position in co-authorship networks in three fields

Artículo publicado en Journal of Informetrics, *9*, 135-144. doi: 10.1016/j.joi.2014.12.001. *Autores: María Bordons,* Javier Aparicio, Borja González-Albo, Adrián A. Díaz-Faes.

Abstract: Research networks play a crucial role in the production of new knowledge since collaboration contributes to determine the cognitive and social structure of scientific fields and has a positive influence on research. This paper analyses the structure of co-authorship networks in three different fields (Nanoscience, Pharmacology and Statistics) in Spain over a three-year period (2006–2008) and explores the relationship between the research performance of scientists and their position in co-authorship networks. A denser co-authorship network is found in the two experimental fields than in Statistics, where the network is of a less connected and more fragmented nature. Using the g-index as a proxy for individual research performance, a Poisson regression model is used to explore how performance is related to different co-authorship network measures and to disclose interfield differences. The number of co-authors (degree centrality) and the strength of links show a positive relationship with the g-index in the three fields. Local cohesion presents a negative relationship with the g-index in the two experimental fields, where open networks and the diversity of co-authors seem to be beneficial. No clear advantages from intermediary positions (high betweenness) or from being linked to well-connected authors (high eigenvector) can be inferred from this analysis. In terms of g-index, the benefits derived by authors from their position in co-authorship networks are larger in the two experimental fields than in the theoretical one.

Keywords: Research performance, Collaboration, Social network analysis, Co-authorship, G-index, Poisson regression model.

8.1. Introduction

Science is increasingly becoming a collaborative endeavour. Collaboration allows scientists to share knowledge, expertise and techniques, expedites the research process, and increases visibility (Katz & Martin, 1997 and Sonnenwald, 2007). Under the assumption of the importance and benefits of collaboration for the advancement of science, scientific collaboration is encouraged by policy makers and the collaboration process is the subject of many academic studies.

From a bibliometric standpoint, collaboration is usually analysed through co-authorship in scientific publications. This indicator presents several limitations, since all co-authorships are sometimes not based on collaborative contributions (e.g. honorary authorship) and not all authors who collaborate become co-authors (Laudel, 2002). However, a positive correlation between collaboration and co-authorship has been described in the literature and this indicator has proved useful to study different aspects of collaboration in science (see for example, Glänzel & Schubert, 2004). Accordingly, co-authorship is used as a measure of scientific collaboration in this paper, although we should have in mind its limitations.

Different indicators have been introduced to quantify collaboration in research papers (see for example, Egghe, 1991; Glänzel & Schubert, 2004 and Vinkler, 2010) and extensive literature has been devoted to explore collaboration patterns (Bordons & Gómez, 2000) and the influence of collaboration on the productivity of scientists and on the impact of research (Abramo, D'Angelo & Di Costas, 2009; Bordons, Aparicio & Costas, 2013; Glänzel, 2001 and Lee & Bozeman, 2005). In most recent years, the application of social network analysis to study co-authorship relations has emerged as an interesting approach, since it allows us to visualise and investigate social structures and relations (see for example, Abbasi, Altmann, & Hossain, 2011; Abbasi, Chung, & Hossain, 2012; Jansen, Von Görtz, & Heidler, 2010; Li-Chun, Kretschmer, Hanneman, & Ze-Yuan, 2006; Newman, 2001 and Otte & Rousseau, 2002). Studies of co-authorship networks may focus on the global structure of networks (macro-perspective) (see for example, Newman, 2001), on the study of subsets (clusters or components) formed within the network (meso-perspective) (He, Ding, & Ni, 2011) or on the individual scientists included in the network's membership (micro-perspective) (for example, Hou, Kretschmer, & Liu, 2008).

Different studies suggest that research networks play a crucial role in the production of new knowledge. The basic idea is that "the position of a node in a network determines in part the opportunities and constraints that it encounters, and in this way plays an important role in a node's outcomes" (Borgatti, Mehra, Brass, & Labianca, 2009). In other words, this means that the position of a scientist in the co-authorship network may have an influence on his/her research performance. This is clearly related to the notion of "social capital", defined as the benefits that actors derive from their social relationships (Coleman, 1988), which may contribute to knowledge creation and to human capital development (Liao, 2011). Three different dimensions of social capital have been described (Nahapiet & Ghoshal, 1998), namely, cognitive capital, relational capital, and structural capital. The latter is the main subject-matter of this study and it can be defined as the value or advantage accrued by an individual or group arising from the structure of social relationships.

There is no consensus on which type of network structure performs best. According to Coleman (1988), densely embedded closed networks are advantageous because they foster the building of mutual confidence and partners bind themselves to one another through reciprocal obligations and expectations ("closure argument"). On the other hand, an alternative view considers that social structural advantages derive from the brokerage opportunities created by an open social structure (Burt, 1992 and Burt, 2004), since it fosters the flow of knowledge between heterogeneous actors and reduces redundant contacts. From this perspective, separate groups control different information and resources, and individuals who bring together people from the different groups act as "brokers" that bridge the existing gaps or "structural holes"[23] between groups ("structural hole argument"). Interestingly, these two notions of social capital are not necessarily contradictory, since different network structures may generate social capital depending on the purpose of the network and the members involved (Ahuja, 2000 and Klenk, Hickey, & MacLellan, 2010).

The relationship between the position of authors in collaboration networks and their performance, as measured by the number of publications, the number of citations and/or the h-index or the g-index, as the case may be, has been previously analysed in the literature. A positive correlation between different centrality measures and citation counts has been described in the fields of information systems (Liao, 2011) and library and information science (Yan & Ding, 2009), while centrality measures showed a positive correlation with scientific output in scientometrics (Hou et al., 2008); these results suggesting that researchers with a higher number of collaborators (high degree) or those who are close to all others in the network (high closeness) are likely to obtain better performance results. Moreover, the influence of the strength of the ties among authors has attracted considerable attention in a number of studies. Scholars who have strong ties (repeated co-authorships) to co-authors earned better research performance results than those with weak ties (single co-authorships with many different authors) in a study on information science (Abbasi et al., 2011). In this study, having an efficient network, with a low rate of redundant contacts, enhanced research performance probably because redundant contacts are less frequently associated with groundbreaking initiatives since they do not provide access to new information. Conversely, establishing connections with researchers in new and diverse teams, bridging structural holes, appeared to be positive for research performance. A positive effect of structural holes on a researcher's performance, as measured by citation scores and individual creativity, was described also in a study in nanoscience (Heinze & Bauer, 2007), while the development of closed social networks with strong ties was positive in other studies on the biotechnology (Walker, Kogut, & Shan, 1997) and pharmaceutical industries (Guler & Nerkar, 2012). As mentioned above, the effect

[23] A structural hole is the absence of ties among a pair of nodes in the ego network (Burt, 1992). The ego is the individual, team or organisational unit under analysis.

of structural holes on performance may vary depending on the context and the field. In this sense, Ahuja (2000) suggests that closed networks are beneficial when strong collaboration is required, while structural holes are likely to be more advantageous when access to diverse information is essential. On the other hand, the positive effect of structural holes may be higher in new fields (such as nanoscience) where brokerage positions become particularly significant because diverse knowledge and ideas are essential for the development of the field.

The patterns and consequences of network structures on scientific or innovative results have been studied in the literature at different levels of analysis, which range from individual scientists (Hou et al., 2008, Klenk et al., 2010 and Li-Chun et al., 2006) or teams (Reagans & Zuckerman, 2001), to higher organisational units such as firms (Ahuja, 2000 and Guler & Nerkar, 2012). Most of these studies deal with the analysis of publications on a given topic or field, whereas interfield comparisons are less frequently addressed. Special mention must be made of the study by Jansen et al. (2010) on the fields of astrophysics and nanoscience concluding that the relationship between network structure and the production of new knowledge is field specific, probably because fields differ in their cognitive structure and knowledge production dynamics.

The objective of this paper is to study the co-authorship networks existing in three different fields (macro-perspective) and explore the relationship between social network measures and research performance of authors with special emphasis on interfield differences (micro-perspective). The assumption that the experimental/theoretical character of a field and its degree of interdisciplinarity may have an influence on its cognitive and social structure was used as the main driver for the selection of our fields of study: one experimental and well-established field (Pharmacology), one experimental, emergent and interdisciplinary field (Nanoscience), and a theoretical field (Statistics).

The interest of this type of study is manifold. The analysis of the fields' structure through the study of their co-authorship networks and the examination of the relationship between social network measures and the research performance of authors may enable us to understand knowledge production dynamics in each field, to figure out which practices are linked to higher performance results and to identify the authors who have a more strategic position within the networks.

8.2. Research question

The questions addressed in this study are as follows:

- ✓ What are the main differences in the structure of fields according to social network measures based in co-authorship analysis?
- ✓ Is there any relationship between the position of a scientist within his/her co-authorship network and his/her research performance? If so, which of the social network-based measures shows a stronger relationship with the performance of scientists? Are there any interfield differences?

8.3. Methods

Scientific publications of Spain on Statistics/Probability, Pharmacy/Pharmacology and Nanoscience/Nanotechnology over the 2006-2008 period were downloaded from the Web of Science database (Science Citation Index Expanded, Social Science Citation Index and Arts & Humanities Citation Index). Disciplines were defined according to the classification of journals into subfields described by the Web of Science.

To cope with the inconsistencies in the names of authors, we used different algorithms aimed at the normalisation of names. These take into account text similarity of names, number of collaborators in common, number of publication journals in common and author subfield to identify pairs of names that are likely to correspond to the same author (Costas & Bordons, 2007).

A matrix including co-authorship frequencies was built for the social network analysis, while different research performance measures were calculated for individual scientists.

8.3.1. Social network measures

Social networks are usually represented by graphs, which include nodes and links. In this paper, nodes correspond to authors and links represent the cooperation relationship between authors on a joint publication. The Pajek software (Batagelj & Mrvar, 2013) was used to graph the network (not shown in this paper) and to calculate the network measures, which can be grouped in two different types: (a) centrality measures and (b) measures of cohesion.

Centrality measures

Centrality measures are useful to analyse how "central" an individual node is to a network. Different measures have been described:

- ✓ *Degree centrality*: This is the number of other nodes connected directly to a given node; therefore, in a co-authorship network, the degree of an author is the number of his/her different co-authors. It is a measure of local centrality (Scott,

1991). Since our interest was to compare node centrality across fields, which have networks with different sizes, a standardised value was calculated. The standardised degree centrality normalises the actual number of links by the maximum number of links it could have (Freeman, 1979), that is: normalised degree = degree/(n − 1), where n is the number of nodes in the network. The normalised degree ranges from 0 (isolated node) to 1 (if the node is connected to all others).

- ✓ *Closeness centrality*: A node is globally central if it lies in average at the shortest distance from all other nodes. It focuses on "how close" an actor is to all other actors in the network (Freeman, 1979). Degree centrality identifies actors who are locally influential (it takes into account the immediate links that a node has), but closeness centrality focuses on the influence of a node over the entire network. An standardised value was calculated to make inter-field comparisons possible: normalised closeness = closeness/(n − 1). The normalised closeness ranges from 0 to 1. This index is only meaningful for a connected network, so it was only applied to the main component.

- ✓ *Betweenness centrality*: The betweenness centrality of a vertex in a graph is calculated as the number of geodesics passing through that vertex. A geodesic is the shortest path between two vertices. In a connected, undirected graph with n vertices, there are at least $n(n-1)$ geodesics. The betweenness centrality can be normalised using $(n-1)(n-2)/2$, which is the maximum number of shortest paths (excluding the node under consideration) (Abbasi et al., 2011). The normalised betweenness = betweenness/$[(n-1)(n-2)]/2 = 2*$betweenness/$(n^2 - 3n + 2)$. It ranges from 0 (a node lies on all geodesics of all pairs of nodes) to 1 (a node lies on no geodesic). In social networks, actors with high betweenness represent gatekeepers or information brokers because they lie among many paths of information flow.

- ✓ *Centralisation:* Centrality measures characterise an actor's position in a network (micro-level measure), while centralisation characterises the whole network (macro-level measure). It indicates how unequal the distribution of centrality is in a network. Degree centralisation in a network is calculated as the variation in the degrees of vertices divided by the maximum degree which is possible in the network of the same size (Wasserman & Faust, 1994). Networks where one or a few nodes show much higher centrality than the other nodes are highly centralised while those in which centrality measures do not differ significantly among nodes show low centralisation. It ranges from 0 (low centralisation) to 1 (high centralisation). In the same way, the betweenness centralisation was calculated.

Closeness centralisation is not shown because it is meaningful only for connected networks.

- ✓ *Eigenvector centrality:* This takes into account not only the number of adjacent nodes but also the values of centrality of these adjacent nodes assuming that a node which is connected to many other nodes that are themselves well-connected has a high eigenvector centrality. Kleimberg (1999) method is used, that is close to Bonacich power (Bonacich, 1972).

Measures of cohesion

Various measures related to the structural cohesion of the networks were considered.

- ✓ *Strength of ties*: The strength of a tie between node *i* and *j* is the weight of the link w_{ij} between those nodes. The weight is the number of co-authorships between two scholars. To assess a node ties strength we obtained the average of the weights of his co-authorships, that is, the number of co-authorships divided by the node degree (Abbasi et al., 2011).

- ✓ *Network constraint*: This allows assessing whether the research networks of the research groups are concentrated directly or indirectly on a single contact (which means no access to structural holes) (Burt, 1992), that is, it allows us to measure how open or closed research networks are. It can be calculated as follows:

$$C_{ij} = (p_{ij} + \sum_q p_{iq}\, p_{qj})^2, \text{ for } q \neq i, j.$$

P_{ij} is the proportion of *i*'s relations directly invested in connection with *j.* The next figure in brackets is the proportion of *i*'s relations that are indirectly invested in connection with contact *j* (Burt, 2004). Constraint is a measure of redundancy of contacts. If an individual's contacts are highly connected to each other, he/she has many redundant contacts and his/her network is highly constrained (Abbasi et al., 2012).

- ✓ *Clustering coefficient*: This is the average of the densities of the neighbourhoods of all actors in a network. It measures to what extent each actor in a network is "embedded" in a local cluster. It is the probability that two neighbours of a vertex are adjacent to each other, that is, the probability that two of a scientist's collaborator have themselves collaborated (Abbasi et al., 2011 and Barabási et al., 2002). A low clustering coefficient for an author means that his/her non-connected co-authors have low probability of writing a joint paper. This is the measure which

provides more specific information about cohesion. When this measure in a network is high, all actors are embedded in cohesive local neighbourhoods (Hanneman & Riddle, 2005).

8.3.2. Measures of research performance

The following indicators were calculated for each author.

- ✓ *Number of articles*: number of articles published in journals covered by the Web of Science database (WoS) (Science Citation Index Expanded, Social Sciences Citation Index, and Arts & Humanities Citation Index). Only articles, reviews and proceedings papers were considered.

- ✓ *Total number of citations* received by the articles in WoS journals. Citations from publication year to February 2014 were counted.

- ✓ *Number of citations per article*: This is the average number of citations received by articles published by a given scientist.

- ✓ *G-index*: Given a set of articles ranked in descending order of the number of citations received, the g-index is the (unique) highest number so that the top *g* articles received (altogether) at least g^2 citations (Egghe, 2006). The advantage of the g-index is that it measures quantity and impact of research by means of a single indicator. The g-index was introduced in 2006 as an improvement to Hirsch's h-index (Hirsch, 2005) because it takes into account the citation scores of top articles and this yields a more precise distinction between scientists from the point of view of visibility.

A regression analysis was used to explore to what extent there is a relationship between the g-index of scientists and their position in the social networks (social network measures as explanatory variables). Since the g-index tends to approximate the form of a Poisson distribution (it takes only positive integer values, it exhibits a positive skew, and the mean and the variance show very similar results), the Poisson multiple regression model was retained. In addition, the Kruskal Wallis test was applied to compare g-index distribution (which does not comply with the normal distribution assumption) between multiple groups of authors. The α level was fixed at 5%. Statistical analyses were conducted with SPSS (version 19)[24].

[24] As an alternative regression model, the relationship between the variable No. Publications*No. Citations (dependent variable) and social network based measures (explanatory variables) was also explored using a negative binomial regression. The results obtained were very similar to those derived from the g-index based regression, but lower R^2 values were achieved, so only the first model is shown.

8.4. Results

The scientific output of Spain for the 2006–2008 period amounted to 943 articles in Statistics/Probability; 1087, in Nanoscience/Nanotechnology; and 2858 in Pharmacology/Pharmacy. The total number of authors in each field and the resulting productivity per author are shown in the first panel of Table 8.1. Higher average team size is observed in Nanoscience and Pharmacology than in Statistics in accordance with their higher co-authorship index (CI), which can be accounted for by a stronger need for collaboration in the experimental fields. An in-depth study of co-authorship links is further conducted through social network analysis.

Table 8.1. General description of the networks (macro-level).

	Statistics	**Nanoscience**	**Pharmacology**
Total network			
No.Articles	943	1,087	2,858
No. Authors	1,572	3,505	10,099
N.Art./author	0.60	0.31	0.28
No.Authors/art (CI)	1.67	3.22	3.53
Reduced network*			
No.Authors	429	1,013	2,609
No.Edges	603	3,106	9,410
Degree centralisation	0.033	0.038	0.025
Betweenness centralisation	0.049	0.165	0.044
No.Components	80	75	162
No.Authors in main component (%)	119 (27.74%)	609 (60.12%)	1,731 (66.35%)
Mean distance	5.34	7.63	8.04
Largest distance	14	23	22

*Reduced network: only non-isolated authors (degree>0) with more than 1 article are considered.

8.4.1. Network structure

A general description of the networks is shown, first at the macro-level to depict the structure of the entire network (second panel of Table 8.1), and then at the micro-level through different measures that characterise the behaviour of authors on the basis of their relationships with other authors (Table 8.2). This study focuses on the set of non-isolated authors (degree > 0) with more than 1 article which constitutes what we have termed as the “reduced network” and is the subject of study in this research paper.

Table 8.2. Structural network measures of authors (micro-level).

	Statistics (n=429)		**Nanoscience** (n=1,013)		**Pharmacology** (n=2,609)	
	Av	SD	Av	SD	Av	SD
Degree	2.81	1.90	6.13	4.51	7.21	5.52
Std_Degree	.007	.004	.006	.004	.003	.002
Std_Closeness	.021	.020	.052	.039	.058	.041
Std_Betweenness	.001	.004	.002	.009	.001	.004
Eigenvector	.005	.048	.004	.031	.001	.020
Clustering coefficient	.611	.431	.776	.300	.775	.277
Constraint	.843	.246	.596	.256	.542	.255
Strength	2.10	0.88	1.85	0.57	1.83	0.76

Note: Av = average; SD = Standard deviation; Std = standarised.
Reduced network: only non-isolated authors (degree>0) with more than 1 article are considered.

There are several differences among the three fields worth pointing out. Firstly, a dense network is observed in Nanoscience and Pharmacology, where the number of lines is far higher than the number of vertices, whereas the network in Statistics may be qualified as sparse (De Nooy, Mrvar, & Batagelj, 2005) since the number of lines in the graph is of the same order as the number of vertices. Secondly, the networks show low values of centralisation in all three fields, that is, centrality is not concentrated in a low number of nodes. Anyway, Nanoscience, if any, is the field which shows the highest centralisation, especially concerning betweenness, because a few authors show relatively high betweenness values. Thirdly, the main component includes around two thirds of the authors in the denser networks (Pharmacology and Nanoscience), as against only 28% in Statistics, which shows a more fragmented structure. One of the underlying reasons for this divergence rests with the fact that collaboration is essential in experimental fields, such as Pharmacology and Nanoscience, where laboratory teamwork is essential; while it is not so indispensable in theoretical fields such as Statistics, where scientists are more likely to work alone or in small teams.

Table 8.2 shows the summary statistics of the structural network measures of authors in the three fields under analysis. Pharmacology and Nanoscience are quite similar according to the patterns of relationships of their authors, while Statistics shows a closer network (higher constraint), stronger links between authors (higher strength) and weaker local cohesion according to the lower propensity of authors to form cliques (lower clustering coefficient). The high number of articles with 2–3 authors in Statistics (68% vs. 20% in Pharmacology and Nanotechnology) contributes to explain its

higher constraint, since authors with a high number of collaborators (high degree) are more likely to have non-redundant contacts. On the other hand, the high number of authors with only one co-author in Statistics (22% of authors vs. 4–6% in Pharmacology and Nanoscience, respectively) contributes to explain the lower propensity of authors to form cliques in that field, since at least two co-authors are needed to form a clique[25].

Summary statistics of the performance of authors by field are shown in Table 8.3. Although outstanding interfield differences in the average citedness of authors' papers or in the g-index of scientists are observed, they cannot be compared due to differences in publication and citation practices by field (Moed, 2005).

Table 8.3. Summary statistics of the performance of authors by field.

	Statistics	**Nanoscience**	**Pharmacology**
N.Authors	429	1,013	2,609
N.Art./author	3.49 (2.6)	3.11 (2.01)	3.29 (3.05)
No.Cit/art.	1.38 (2.49)	21.34 (23.38)	18.86 (17.95)
G-index	1.21 (1.02)	2.96 (1.78)	3.08 (2.10)

8.4.2. Relationship between performance indicators and the position of authors in networks

To explore to what extent changes in the co-authorship network measures contribute to explain changes in the g-index we used a Poisson regression model. The predictor variables introduced include seven continuous ones for each of the fields: standardised betweenness, standardised closeness, standardised degree, eigenvector, clustering coefficient, average ties strength and constraint. Unfortunately, constraint had to be removed from the analysis due to multicolinearity problems. To allow for the comparison of variables which are expressed in different units of measurement, continuous variables are transformed to new variables with a mean of 0 and a standard deviation of 1 (*Z*-scores). *Z*-scores are a unit free measure which can be used to compare observations measured with different units. Three different models are built, one for each field, to identify interfield differences in the association between the co-authorship measures and the g-index.

[25] If degree = 1, the clustering coefficient = 0, since no cliques can be formed.

Our results show that the models fit reasonably well. The omnibus test, which compares the fitted model against the intercept-only model, is statistically significant in all three fields ($p < 0.001$) suggesting that changes in the predictor variables contribute to explain changes in the dependent variable. The results of the Poisson regression model are shown in Table 8.4. As a measure of the goodness of fit of the models the correlations between observed and predicted values of g-index are calculated. The best fit is obtained in Pharmacology ($R^2 = 0.652$), followed by Nanoscience ($R^2 = 0.573$) and Statistics ($R^2 = 0.195$).

Table 8.4. Poisson regression analysis for the g-index.

	Statistics				**Nanoscience**				**Pharnacology**			
	β	Hypothesis test		Exp(*β*)	*β*	Hypothesis test		Exp(*β*)	*β*	Hypothesis test		Exp(*β*)
		Wald Chi Square	Sig.			Wald Chi Square	Sig.			Wald Chi Square	Sig.	
(Intercept)	.147	10.38	**.001**	1.159	1.014	2,744.62	**0.000**	2.756	1.042	7,713.91	**0.000**	2.835
ZStd_degree	.277	32.11	.000	1.319	.200	103.19	**0.000**	1.221	.157	256.84	**0.000**	1.170
Zclosenness	-.092	2.75	.097	.912	.032	1.94	.164	1.032	.041	10.03	**.002**	1.042
Zbetweenness	-.050	1.20	.274	.951	-.021	2.25	.134	.979	.013	1.69	.193	1.013
Zclust_coefficient	-.069	2.15	.143	.933	-.167	85.17	**0.000**	.846	-.169	204.24	**0.000**	.844
Zav_strength	.192	23.60	**.000**	1.211	.198	123.44	**0.000**	1.219	.255	939.69	**0.000**	1.291
Zeigenvector	-.022	0.45	.505	.978	.014	1.44	.231	1.014	-.058	74.09	**0.000**	.944
(Scale)	1[a]				1[a]				1[a]			

[a] Fixed at the displayed value.

Even though the interpretation of the coefficients (*β*) in the model may seem difficult due to the nature of the log link function $y = \exp(a + bx)$, a Poisson regression models the log of the expected g-index as a function of the predictor variables. The signs of the coefficients show whether the predictors have a positive or negative association with the g-index. A positive coefficient for a continuous variable indicates a positive relationship between the predictor and the g-index, while a negative coefficient indicates an inverse relationship. Our results suggest a lower association between the g-index and network-based measures in the field of Statistics, since only two variables show a statistically significant relationship with the g-index (as against three variables in Nanoscience and five in Pharmacology) and the goodness of the fit of the final model is weaker in this field.

Degree and tie strength are the variables which display the strongest relationship with the g-index in all three fields under study. Our results suggest that the degree is the most influential variable in Statistics (32% increase in g-index for every unit increase in

degree) while both degree and tie strength show similar influence in Nanoscience (around 22% increase in g-index) and tie strength is the most influential variable in Pharmacology (29% increase in g-index).

Local cohesion as measured by the clustering coefficient shows a negative association with the g-index both in Nanoscience and Pharmacology. This suggests that establishing collaborations with scientists who do not collaborate between them (for example, if they belong to different teams or work in different research lines within a team) is on average positive for the research performance of a given author. On the other hand, considering authors with the highest propensity to form cliques (clustering coefficient = 1), those in Statistics were more likely to obtain a high degree and a high g-index than those in the two experimental fields, a result that would suggest the lower negative effect of close networks on research performance in the theoretical field.

Closeness and eigenvector variables are significant only in Pharmacology. The positive association between closeness and the g-index suggests that global centrality is on average positive for research performance in the field, probably increasing the opportunity to find new collaborators. From a detailed examination of our data we observed that in Pharmacology the highest values of closeness centrality correspond to several scientists affiliated to hospitals who collaborate within their institution but also with scientists in other hospitals as well as in universities and pharmaceutical companies. This position, close to many other actors in the network, allows them to participate in highly relevant pharmacological research, such as that conducted in the framework of clinical trials and, in the long term, to obtain high g-index values.

The negative relationship between the eigenvector and the g-index observed in Pharmacology is somewhat counterintuitive, since better performance could be expected for the best connected authors. In fact, a positive bivariate correlation between the g-index and the eigenvector was observed (Spearman's rho = 0.218). To explain this, we should keep in mind that the beta values in the multiple regression need to be understood in the context of the overall model. Other variables in the model may account for part of the information provided by the eigenvector in such a way that the negative beta value of the eigenvector corrects for an excessive positive influence of another related variable/s. In fact, the negative sign of the eigenvector implies a higher reduction of the g-index for those authors with the highest eigenvector values. This is consistent with our data, since the authors with the highest eigenvector values do not present the highest g-index values. The reason is that a high degree is very relevant to obtain a high g-index and for authors with a very high number of collaborators, it can be difficult to obtain a high eigenvector value, since it is unlikely for all the co-authors of a given author to be well-connected ones.

To gain further insight into the comparative importance of the number of collaborators and the strength of ties on the research performance of scientists, four categories of scientists were distinguished according to their values of degree (high or low) and strength of ties (high or low)[26]. We observe that the g-index differs based on the four-group classification of scientists (Kruskal–Wallis test < 0.001). Figure 8.1 shows that the g-index tends to increase from authors with low degree-low strength (group 1) to those with high degree-high strength (group 4) in the three fields. Cumulative positive effects of degree and strength of ties can be observed in the fourth category. However, only in Nanoscience the differences between the g-index of the two intermediate categories are statistically significant ($p < 0.001$): g-index values tend to be higher for authors with high degree and low strength compared to those with low degree and high strength, thus suggesting that a diversity of links may outweigh the negative effect of low strength in that field.

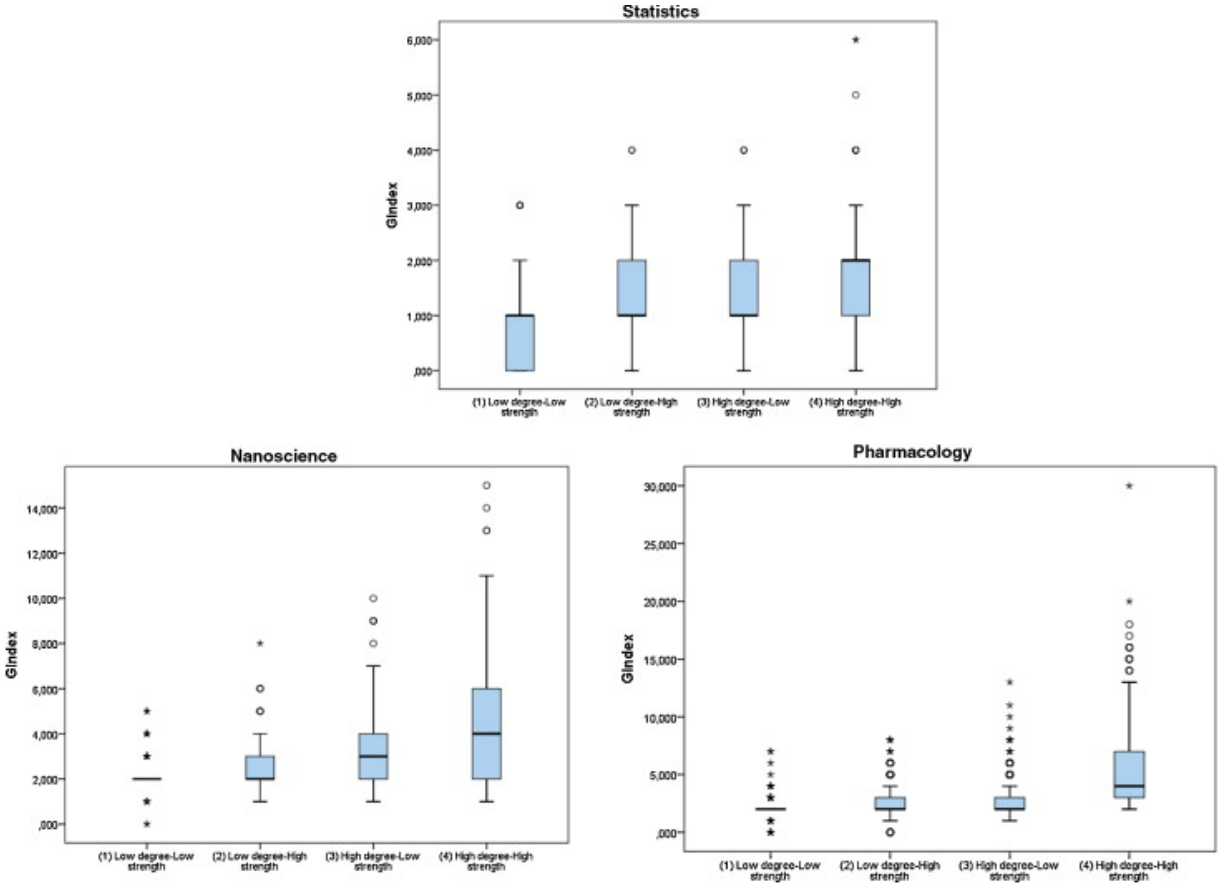


Figure 8.1. G-index by level of degree and strength of ties for authors in the three fields.

Note: significant differences between all pair of groups in Nanoscience (p<0.01); between all pairs of groups (p<0.001) except 2 vs 3 in Pharmacology; between groups 1 vs.3 and 1 vs.4 in Statistics (p<0.01).

[26] The median is considered in each field to separate "low" values (≤median) from "high" values (>median).

8.5. Discussion

This study shows differences in the structure of the collaboration networks in the three fields under analysis. At the macro level, Pharmacology and Nanoscience present a similar network structure, denser than Statistics, which displays a less connected and more fragmented network. The underlying reason is the bigger size of teams in laboratory-based research conducted in experimental fields such as Pharmacology and Nanoscience when compared to Statistics, where scientists are more likely to work alone or in small teams.

The study at the micro-level confirms that there is a relationship between the position of Spanish scientists in co-authorship networks and his/her research performance as measured by the g-index. This association varies by field and seems to be stronger in Pharmacology and Nanoscience than in Statistics.

In all three fields under study, the variables which show a stronger relationship with the g-index are the average strength and the standardised degree. Specifically, scientists who have many collaborations with different scholars (high degree) or that build strong links with their co-authors (high average strength) are more likely to show a higher g-index. Among the benefits of a higher number of direct ties (high degree), knowledge sharing through interaction and discussion can be mentioned. Scientists can learn from one another and produce better research if they pool their knowledge, skills and resources (see for example, Abbasi et al., 2011 and Badar, Hite, & Badir, 2013). On the other hand, repeated co-authorships may be accounted for by mutual confidence and a set of shared norms of behaviour between the partners, which can facilitate resource sharing and cooperation (see for example, Abbasi et al., 2011, Ahuja, 2000 and Guler & Nerkar, 2012). In our study, there is a positive association between the g-index and both the number of collaborators and the strength of links, but the diversity of co-authors seems to be more important in Statistics, while the strength of the relationships with existing co-authors seems to be more relevant in Pharmacology. Strong ties can be more relevant in experimental fields such as Pharmacology due to the closer interaction and reciprocal support among members needed to conduct laboratory work. This is probably not so evident in Nanoscience – in spite of its experimental nature – because of the higher importance of diversity of sources (degree) in the more interdisciplinary fields (Jansen et al., 2010). In other fields, such as Information Systems and Information Technology, expanding social relationships, especially with different co-authors (Abbasi et al., 2011) but also with the same co-authors (Liao, 2011), also emerged as an effective way to improve research performance. Interestingly, authors with a high number of collaborators and strong ties show in our study the highest g-index values across all fields, although this was not the most common situation. Comparing the g-index of authors with a high

number of collaborators but low tie strength values with those showing a low number of collaborators but high tie strength values, significant differences were only found in Nanoscience, confirming the higher benefit drawn by the first set of authors. This finding is consistent with the important role played by the diversity of links in the more interdisciplinary fields above mentioned.

As regards the ongoing debate about which type of network (closed or open) is more beneficial for performance, we were not able to approach the subject through the study of constraint, a common feature in the literature, since this variable was removed from the analysis due to multicolinearity problems. However, our data point to a negative association between g-index and local cohesion (clustering coefficient) in the experimental fields, which means that widening the network of collaborators to scientists who do not collaborate between them is on average positive, at least in Pharmacology and Nanoscience, thereby suggesting that more open structures would be more beneficial in these fields.

Being a well-connected author (as measured by the eigenvector) is not associated to a higher g-index in two fields, while a negative association is observed in Pharmacology. An inverse relationship between research performance and the eigenvector was also reported by Abbasi et al. (2011) in a study on social networks in Information Science. The fact that well-performing scientists (i.e. research leaders of teams) had a great proportion of their papers written in collaboration with students rather than with other well-performing scientists was the explanatory reason held for this. In our study, we have observed that it can be especially difficult for scientists with a very high g-index to obtain a high eigenvector value, since they usually have many collaborators and it is unlikely for all their co-authors to be well-connected. The essential role of teams in Pharmacology, which include members in different stages of their scientific career (from students to senior scientists) and with different levels of productivity, needs to be considered to understand that collaboration limited to well-connected authors is not the norm in the field even for senior scientists, who maintain links with authors which may differ largely in their structural positions within the network.

The benefits of being, geodesically speaking, between many authors (high betweenness) has been reported in the literature (Li et al., 2013 and Yan & Ding, 2009), and are mainly based on the fact that these scientists have ties connecting otherwise disconnected authors thus enabling access to diverse sources of knowledge. However, our data suggest that in the case of Spanish scientists, playing a bridging role is not associated to a higher g-index in any of the three fields under analysis. As stated by Abbasi et al. (2011), “brokerage positions” may have strategic value, but do not necessarily improve research performance, maybe due to the costs of maintaining collaboration with authors from different contexts.

In summary, our study shows there is a relationship between the position of scientists in the co-authorship network and their research performance, with these relationships being stronger in the experimental fields (Pharmacology and Nanoscience) than in Statistics. Having a high number of collaborators and/or high strength of links with co-authors is associated to a higher g-index of scientists in all three fields. Including collaborators from different contexts, who do not collaborate between them, is also found to be a positive factor in all fields with the exception of Statistics. Being close to all other authors in the network is significant in Pharmacology, because these central positions are occupied by scientists who connect teams from different institutions and participate in highly relevant and collaborative research. No clear benefits from intermediary positions (high betweenness) or from those in connection with well-connected authors (high eigenvector centrality) are derived from this study.

Our research is subject to a series of limitations. (1) Firstly, our study is based on co-authorship, which is a partial indicator of scientific collaboration (Katz & Martin, 1997), so not all collaborative links between scientists are considered (e.g. those collaborators mentioned in the acknowledgment section or not mentioned at all are not visible). (2) We have used a single measure to assess research performance which indeed is a multidimensional endeavour and would require more complex measures. Moreover, the specific limitations described in the literature for the g-index, such as being affected by an occasional “big hit” (a highly cited document) (Costas & Bordons, 2008), apply to our study as well. (3) Conclusions derived from the association between the g-index of the authors and their network-based measures should be read with caution, since regression analyses reveal relationship among variables but do not imply that the relationships are causal, and the influence of other unmeasured variables cannot be discarded. (4) We do not know to what extent the delimitation of the fields according to WoS subject categories could influence the results, for example underestimating the interdisciplinarity of the fields. (5) Our results describe the behaviour of Spanish authors in three different fields, and may not be extrapolated to other communities of scientists.

Acknowledgments

This research was supported by the Spanish Ministry of Science and Innovation (MICINN) (research project CSO2008-06310) and the Spanish National Research Council (JAE predoctoral grant and project 201110E087). We are very grateful to Laura Barrios and José Manuel Rojo for their statistical advice as well as to two anonymous referees for their valuable comments.

References

Abbasi, A., Altmann, J., & Hossain, L. (2011). Identifying the effects of co-authorsip networks on the performance of scholars: a correlation and regression analysis of performance measures and social network analysis measures. *Journal of Informetrics*, *5*, 594-607.

Abbasi, A., Chung, K.S.K., & Hossain, L. (2012). Egocentric analysis of coauthorship network structure, position and performance. *Information Processing and Management*, *48*(4), 671-679.

Abramo, G., D'Angelo, C.A., & Di Costa, F. (2009). Research collaboration and productivity: is there correlation? *Higher Education*, *57*, 155-171.

Ahuja, G. (2000). Collaboration networks, structural holes and innovation: A longitudinal study. *Administrative Science Quarterly*, *45*(3), 425-455.

Badar, K., Hite, J.M., Badir, & Y.F. (2013). Examining the relationship of coauthorship network centrality and gender on academic research performance: the case of chemistry researchers in Pakinstan. *Scientometrics*, *94*, 755-775.

Barabási, A. L., Jeong, H., Neda, Z., Ravasz, E., Schubert, A., & Vicsek, T. (2002). Evolution of the social network of scientific collaborations. *Physica a-Statistical Mechanics and Its Applications, 311*(3-4), 590-614.

Batagelj, V., Mrvar, A. (2013). Pajek. V 3.14. http://vlado.fmf.uni-lj.si/pub/networks/pajek/.

Bonacich, P. (1972). Factoring and weighting approaches to status scores and clique identification *Journal of Mathematical Sociology*, 2(1), 113-120.

Bordons, M., Aparicio, J., & Costas, R. (2013). Heterogeneity of collaboration and its relationship with research impact in a biomedical field. *Scientometrics*, *96* (2): 443-466.

Bordons, M., & Gómez, I. (2000). Collaboration networks in science. In: B.Cronin and H.B.Atkins (Eds.) The web of knowledge: A festschrift in honor of Eugene Garfield (pp.197-213). Medford, NJ: Information Today.

Borgatti, S.P., Mehra, A., Brass, D.J., & Labianca, G. (2009). Network analysis in the social sciences. *Science*, *323,* 892-895.

Burt, R.S. (1992). Structural holes. Cambridge, MA: Harvard University Press.

Burt, R.S. (2004). Structural holes and good ideas. *American Journal of Sociology, 110*(2), 349-399.

Coleman, J.S. (1988). Social capital in the creation of human capital. American *Journal of Sociology*, *94*(Supplement), 95-120.

Costas, R., & Bordons, M. (2007). Algorithms to solve the lack of normalization in author names in bibliometric studies. *Investigación bibliotecológica*, *21*(42), 13-32.

Costas, R., & Bordons, M. (2008). Is g-index better than h-index? An exploratory study at the individual level. *Scientometrics*, *77*(2), 267-288.

De Nooy, W., Mrvar, A., & Batagelj, V (2005). Exploratory network analysis with Pajek. Cambridge: Cambridge University Press.

Egghe, L. (1991). Theory of collaboration and colaborative measures. *Information Processing and Management*, *27*, 177-202.

Egghe, L. (2006). Theory and practise of the g-index. *Scientometrics*, *69*(1), 131–152.

Freeman, L.C. (1979). Centrality in social networks: Conceptual clarification. *Social Networks*, *1*(3), 215-239.

Glänzel, W. (2001). National characteristics in international scientific co-authorship relations. *Scientometrics*, *51*(1), 69-115.

Glänzel, W., Schubert, A. (2004). Analysing scientific networks trough co-authorship. In: Moed, H.F., Glänzel, W. and Schmoch, U. Ed. Handbook of Quantitative S&T Research. Dordrecht: Kluwer Academic Publisher, pp. 257-276.

Guler, I., & Nerkar, A. (2012). The impact of global and local cohesion on innovation in the pharmaceutical industry. *Strategic Management Journal*, *33*, 535-549.

Hanneman, R.A., & Riddle, M. (2005). Introduction to social network methods. University of California at Riverside: Riverside, CA.

He, B., Ding, Y., & Ni, C. (2011). Mining enrich contextual information of scientific collaboration: a meso perspective. *Journal of the American Society for Information Science and Technology*, *62*(5), 831-845.

Heinze, T., & Bauer, G. (2007). Characterizig creative scientists in nano-S&T: Productivity, multidisciplinarity, and network brokerage in a longitudinal perspective. *Scientometrics,* 70(3), 81-830.

Hirsch, J.E. (2005). An index to quantify and individual's scientific research output. *Proceedings of the National Academy of Sciences*, *102*(46), 16569–16572.

Hou, H., Kretschmer, H., & Liu, Z. (2008). The structure of scientific collaboration networks in Scientometrics. *Scientometrics, 75*(2), 189-202.

Jansen, D., Von Görtz, R., & Heidler, R. (2010). Knowledge production and the structure of collaboration netwoks in two scientific fields. *Scientometrics,* 83(1), 219-241.

Katz, S., & Martin, B.R. (1997). What is research collaboration? *Research Policy*, *26,* 1-18.

Kleinberg, J.M. (1999). Authoritative Sources in a Hyperlinked Environment. *Journal of the ACM*, *46*(5), 604-632.

Klenk, N.L., Hickey, G.M., & MacLellan, J.I. (2010). Evaluating the social capital accrued in large research networks: the case of the Sustainable Forest Management Network (1995-2009). *Social Studies of Science*, *40*(6), 931-960.

Laudel, G. (2002). What do we measure by co-authorships? Research Evaluation 11(1): 3-15.

Lee, S., & Bozeman, B. (2005). The impact of research collaboration on scientific productivity. *Social Studies of Science, 35*, 673-702.

Li, E.Y., Liao, C.H., & Yen, H.R. (2013). Co-authorship networks and research impact: a social capital perspective. *Research Policy*, *42,* 1515-1530.

Liao, C.H. (2011). How to improve research quality? Examining the impacts of collaboration intensity and member diversity in collaboration networks. *Scientometrics, 86*, 747-761.

Li-Chun, Y., Kretschmer, H., Hanneman, R.A., & Ze-Yuan, L. (2006). Connection and stratification in research collaboration: an analysis of the COLLNET network. *Information Processing and Management*, *42*, 1599-1613.

Moed, H. F. (2005). Citation analysis in research evaluation. Dordrecht: Springer.

Nahapiet, J., & Ghoshal, S. (1998). Social capital, intellectual capital and the organizational advantage. *Academy of Management Review*, *23*(2), 242-266.

Newman, M.E.J. (2001). The structure of scientific collaboration networks. *PNAS*, *98*(2), 404-409.

Otte, E., & Rousseau, R. (2002). Social network analysis: A powerful strategy, also for the information sciences. *Journal of Information Science*, *28*, 443-455.

Reagans, R., & Zuckerman, E.W. (2001). Networks, diversity and productivity: the social capital of corporate R&D teams. *Organization Science*, *12*(4), 502-517.

Scott, J. (1991). Social network analysis. A handbook. London: Sage Publications.

Sonnenwald, D.H. (2007). Scientific collaboration. *Annual Review of Information Science and Technology*, *41*(1), 643-681.

Vinkler, P. (2010). The evaluation of research by scientometric indicators.

Oxford: Chandos Publishing.

Walker, G., Kogut, B., & Shan, W. (1997). Social capital, structural holes and the formation of an industry network. *Organization Science*, 8, 109-125.

Wasserman, S., & Faust, K, (1994). Social network analysis: Methods and applications. Cambridge: Cambridge University Press.

Yan, E., & Ding, Y. (2009). Applying centrality measures to impact analysis: a co-authorship network analysis. *Journal of the American Society for Information Science and Technology, 60*(10), 2107-2118.

9. Acknowledgments in scientific publications: presence in Spanish science and text patterns across disciplines

Artículo publicado en Journal of the Association for Information Science and Technology, *65*(9), 1834-1849. doi: 10.1002/asi.23081. *Autores:* Adrián A. Díaz-Faes, *María Bordons.*

Abstract: The acknowledgments in scientific publications are an important feature in the scholarly communication process. This research analyzes funding acknowledgment presence in scientific publications and introduces a novel approach for discovering text patterns by discipline in the acknowledgment section of papers. First, the presence of acknowledgments in 38,257 English-language papers published by Spanish researchers in 2010 is studied by subject area on the basis of the funding acknowledgment information available in the Web of Science database. Funding acknowledgments are present in two thirds of Spanish articles, with significant differences by subject area, number of authors, impact factor of journals, and, in one specific area, basic/applied nature of research. Second, the existence of specific acknowledgment patterns in English-language papers of Spanish researchers in 4 selected subject categories (cardiac and cardiovascular systems, economics, evolutionary biology, and statistics and probability) is explored through a combination of text mining and multivariate analyses. “Peer interactive communication” predominates in the more theoretical or social-oriented fields (statistics and probability, economics), whereas the recognition of technical assistance is more common in experimental research (evolutionary biology), and the mention of potential conflicts of interest emerges forcefully in the clinical field (cardiac and cardiovascular systems). The systematic inclusion of structured data about acknowledgments in journal articles and bibliographic databases would have a positive impact on the study of collaboration practices in science.

Keywords: bibliometrics, text mining.

9.1. Introduction

Research is an increasingly collaborative activity. A series of benefits drawn from collaboration, such as the sharing of knowledge, skills, and techniques; the cross-fertilization of ideas; the division of labor; or the increased motivation derived from human interaction, have been mentioned in the literature. Collaboration enables scientists to address increasingly complex research problems and achieve greater effectiveness in the development of research (Bordons & Gómez, 2000). From a bibliometric point of view, the upward trend in co-authorship rates throughout the past decades in most scientific fields is a reflection of the increasing role of

collaboration in science. In parallel, acknowledgments, which may be considered a measure of "sub-authorship collaboration" (Patel, 1973; Heffner, 1981), have also intensified their presence in scientific publications.

Acknowledgments are an important feature in scholarly communication, because they are used to recognize some special contributions to research that do not qualify for authorship status but may well have a significant bearing on the final results of research (Kassirer & Angell, 1991). As stated by Cronin (1995), an acknowledgment is a voluntary act that follows an implicit code of professional conduct. It has become a constitutive feature of the academic journal article throughout the 20th century as well as a potentially rich source of insight into sub-authorship collaboration in science.

Acknowledgments may be made for different reasons, but they are usually expressions of gratitude concerning different types of support received by researchers. Two general types of support were described by Patel (1973), who distinguishes between technical support, including, among other tasks, collecting data, processing data, operating laboratory machinery, and performing statistical analyses and theoretical support, such as reading, editing, and contributing comments to a draft paper. A more detailed typology was suggested by Cronin and colleagues, who have conducted different studies on the function, frequency, and evolution of acknowledgments in journal papers for various disciplines: information science and library science (Cronin, 1991; Cronin, McKenzie, & Stiffer, 1992); history, philosophy, psychology, and sociology (Cronin, McKenzie, & Rubio, 1993; Cronin, Shaw, & La Barre, 2003); and chemistry (Cronin, Shaw, & La Barre, 2004). As a result of a survey conducted to analyze formal acknowledgments of the papers appearing in JASIS from 1970 to 1990, Cronin identified six acknowledgment categories: paymaster, moral support, dogsbody, technical, prime mover, and trusted assessor (Cronin, 1991, 1995). According to this author, *paymaster* covers the recognition of grants or fellowships; *moral support* shows credit for the provision of access to institutional facilities; *dogsbody* refers to the support from colleagues in routine work such as bibliographic checking, data collection, or analyses; *technical* embraces advice on statistical techniques, computer programming, and comparable tasks; *prime mover* is reserved for individuals who have been influential in stimulating or encouraging the study; and *trusted assessor* is for those who have helped with their ideas, suggestions, or insights to shape the work. In more recent studies on psychology and philosophy (Cronin et al., 2003) and chemistry (Cronin et al., 2004) acknowledgments were classified as moral, financial, editorial, instrumental/technical, and conceptual/cognitive. The conceptual category, which was initially described by McCain (1991, p. 512) as "peer interactive communication" (PIC), is particularly interesting. According to this McCain, PIC includes different contributions such as providing specific information or making specific suggestions, providing critical comments on manuscripts, proffering thanks for advice and discussion, and proffering thanks for inspiration. Because conceptual support

implies an intellectual debt, it is especially relevant, to such an extent that some researchers consider it should be deemed at least as valuable as citations (Edge, 1979; Cronin, McKenzie, Rubio, & Weaver-Wozniak, 1993).

In assessing the research performance of scientists, measures based on productivity (number of publications) and impact (citations) play the most relevant role in the prevailing reward system of science. The need for acknowledgments should also be taken into account in what has been dubbed the "reward triangle," which was put forward by Cronin and Weaver (1995, p. 173), who argued that acknowledgments have a social, cognitive, and instrumental meaning that should be studied. As stated critically by these authors, "the most trivial citation counts for more than the most sincere acknowledgment," but both, citations and acknowledgments, describe interaction and influence and should be studied as important features of the scholarly communication process.

Until recently, it was very difficult to carry out studies on acknowledgments (see Giles & Council, 2004), because this information was not available in bibliographic databases. However, the Web of Science (WoS) has been including funding acknowledgment data since August, 2008. This recent development in the WoS database opens up new possibilities for data mining and the analysis of the information contained in the acknowledgment section of papers (Rigby, 2011).

Acknowledgment data can be used for a variety of purposes in science studies, ranging from the study of the interaction among scientists from a sociological standpoint to their uses in research evaluation and funding policy issues. Because many funding bodies mandate being acknowledged in the papers resulting from the research made with their support, the analysis of the acknowledgment information can be useful to track research output and assess the influence of any funding body or specific grant/research program and to identify the strategic scope of a funding agency (Rigby, 2011, 2013). Some studies have analyzed these trends and identified the most acknowledged entities by category (Giles & Council, 2004; Lewison & Markusova, 2010) or by country (Wang & Shapira, 2011; Wang, Liu, Ding, & Wang, 2012). It has been discussed whether funding income might be an indicator of research quality (Gillett, 1991), because research funding entities generally apply some form of peer review to grant applications. This claim is consistent with the results of Costas and van Leeuwen (2012), who analyzed scientific publications covered by the WoS in 2009 and described a greater impact for publications with funding acknowledgments compared with the remaining papers. The impact of grant-funded research was also found to be greater than that of the rest of research in a study dealing with core journals in library and information science (Zhao, 2010) and in a work focusing on U.S. research (Levitt, 2011). Although some research suggests that the impact of research increases with the

number of funding sources (Lewison & Dawson, 1998), no clear relationship between the two variables has been observed in other studies (Rigby, 2011, 2013).

In recent decades, an increase in the use of acknowledgments has been described for different disciplines such as chemistry (Cronin et al., 2004), psychology, and philosophy (Cronin et al., 2003). Interestingly, differences by discipline in the frequency of acknowledgments (Costas & van Leeuwen, 2012) and in the prevailing type of support acknowledged have been observed (Cronin et al., 2004). Acknowledgments seem to be more frequent in the hard sciences but more elaborated in the humanities and social sciences (Salager-Meyer, Alcaraz-Ariza, Luzardo-Briceño, & Jabbour, 2011). Although financial support emerges as the prevailing type in some disciplines such as chemistry (Cronin et al., 2004) and psychology (Cronin et al., 2003), the conceptual type is the most common in other fields, as in the case of philosophy (Cronin et al., 2003). Moreover, the context of publication (for example, the geographic origin of articles and language) also has an influence on the frequency and content of acknowledgments, which seem to be longer and appear more frequently in Anglo-American journals, maybe because acknowledgments have not yet become such a highly institutionalized practice in the non-Anglo-American context (Salager-Meyer, Alcaraz-Ariza, & Pabón-Bervesí, 2009).

Against this backdrop, our research aims to increase our knowledge about the presence and role of acknowledgments in scientific publications. First, the analysis of funding acknowledgment presence in English-language papers published by Spanish researchers is carried out, with special emphasis on differences by subject area. The study focuses on papers written in English because acknowledgments have to be in English to be captured and processed by *Thomson Reuters*[27]. Second, the existence of specific acknowledgment patterns is explored in four disciplines. Although the inclusion of acknowledgment data in bibliographic databases represents an important step forward, their use remains complicated insofar as they include unstructured information (natural-language text). The study of the content of acknowledgments in previous works is generally addressed by visual inspection and the classification of records according to an acknowledgment typology based on motivation, such as the one described by Cronin (1995). An interesting exception is the paper by Costas and van Leeuwen (2012) in which PIC is identified by means of a search strategy in the funding acknowledgment information available in WoS records. In this study, we introduce a novel approach that explores the usefulness of textual data analysis (Lebart & Salem, 1994; Lebart, Salem, & Bécue, 2000) to identify acknowledgment patterns by discipline.

[27] Information provided by the Technical Support Team of *Thomson Reuters*, June 2013.

9.2. Objectives

The acknowledgment section of WoS papers written in English by Spain-based researchers is analyzed in this study with two different and complementary purposes: first, to increase our knowledge about funding acknowledgment presence by subject area and, second, to discover specific acknowledgment patterns by discipline. With respect to the first objective, a variety of questions, such as the following, is addressed: Are there inter-area differences in the presence of acknowledgments in papers? Is there any evidence establishing the higher quality of funded research? Does funded research include a higher number of authors? Are there differences between basic and applied research in their propensity to acknowledge funding? Do the acknowledgment practices of Spanish researchers resemble those of the international scientific community in their corresponding fields? The second objective of this analysis is to assess the possibility of obtaining acknowledgment patterns by discipline. The acknowledgment section includes natural language text, so textual data analysis and multivariate techniques are used to characterize four disciplines based on the type of information included in the acknowledgment section of papers (analysis of lexical profile). It is interesting to point out that the acknowledgment section of papers includes funding data and also sub-authorship information. Therefore, both types of information contribute to define the final disciplinary pattern. Differences between disciplines in their acknowledgment patterns are expected because of divergences in their cultural norms and funding, instrumentation, and teamwork requirements.

Many studies on collaboration in science use only co-authorship-based indicators to analyze collaborative research. However, Melin and Persson (1996) suggest that when we reduce collaboration to co-authorship we are running the risk of neglecting some collaborative activity. According to Laudel (2002), half of the collaborative research practices are overlooked by the classical bibliometric indicator. In this sense, the inclusion of acknowledgments information in the WoS breaks new ground to study collaboration in science from a wider perspective.

9.3. Data and methods

The methodological aspects of this analysis are organized into three different information blocks. First, a description of data sources and the structure of the acknowledgment information in the WoS database is provided. Then, one different information block is obtained for each of the two approaches adopted for the study of the funding acknowledgment section: analysis of funding acknowledgment presence by subject area in Spanish output and textual analysis of acknowledgments in selected disciplines.

9.3.1. Data sources

Scientific papers published in English by Spain-based researchers in 2010 were downloaded from the WoS database in March, 2012. This study focuses on citable items, which include original articles and reviews (hereafter referred to as *papers*). The WoS database includes three sections of information on funding acknowledgment[28] (FA): funding agency (FO) contains the names of the agencies that support the research; grant number (FG) provides project identification numbers, if any; and funding text (FT) contains the full text included by the authors in the acknowledgment section of the paper. An example of the funding acknowledgment data included in a paper in WoS is shown in Table 9.1.

Table 9.1. Example of WoS funding acknowledgment data.

Funding Agency	**Grant Number**	**Funding Text**
Instituto de Salud Carlos III, Madrid, Spain Lund University Hospital Swedish Heart-Lung Foundation	G03/078	The study was supported by a grant from the Instituto de Salud Carlos III, Madrid, Spain (grant code: G03/078). Two research grants from Lund University Hospital and The Swedish Heart-Lung Foundation are acknowledged. The authors would like to thank A.Smith for the statistical analysis of data.

In this article, we have worked with the information in the FT section of the WoS records, although we shall refer to it as *FA* for the sake of clarity, because it was considered a clearer abbreviation for *funding acknowledgment*. Note that acknowledgments are collected in the WoS only when they include funding information, but for all such entries all acknowledgment types included by the authors (not only those of the funding type) are collected. This means that our results on the presence of acknowledgments refer specifically to funding, but we can explore sub-authorship patterns when data on the latter have been collected along with any funding information. We consider that the set of acknowledged funding records provides a representative substratum for the study of acknowledgment patterns by discipline.

9.3.2. Analysis of FA presence by subject area

The presence of FA in Spanish scientific publications is analyzed for the total country and by subject area. Publications were assigned to disciplines following the WoS's

[28] http://wokinfo.com/products_tools/multidisciplinary/webofscience/fundingsearch/

classification of journals into subject categories. In total, 243 subject categories[29] were grouped into 10 subject areas: agriculture, biology, and environment; biomedicine; chemistry; clinical medicine; engineering and technology; humanities; mathematics; multidisciplinary; physics; and social sciences.

A study of the relationship between FA presence and different variables, namely, the prestige of the publication journal, the degree of collaboration in papers, and the basic versus applied nature of research, was undertaken using the following indicators.

- ✓ *Journal prestige:* Papers published in first-quartile journals (Q1; top 25% journals in the impact factor journal ranking) within each *Journal Citation Reports* subject category were identified to explore whether research published in Q1 journals showed a greater presence of FA.
- ✓ *Collaboration:* The average number of authors per paper depending on whether the research was funded or not was studied.
- ✓ *Basic versus applied nature of research:* A classification of journals into four research levels ranging from 1 (most applied level) to 4 (most basic level) was used. The research level was assigned to individual journals on the basis of both expert review and patterns of journal-to-journal citation, in a way that each journal refers mainly to itself and to other journals in the same level or one level more basic. This classification was described by CHI Research/Computer Horizons Inc. (Noma, 1986; Narin, Pinski, & Gee, 1976), which now operates as ipIQ. The average research level of papers depending on whether the research was funded or not was analyzed.

SPSS v.19 was used for the statistical analysis of data. Differences in the presence of FA by genre were studied by applying the χ^2 test. Mann-Whitney's *U* test for non-parametric distributions was performed to explore differences in the average number of authors and in the average research level of papers according to FA presence (funded vs. nonfunded papers). The relationship between the percentage of papers with FA in total journals and in Q1 journals was studied by subject area and subject category through Spearman's ρ coefficient. The α level was fixed at 5%.

9.3.3. Analysis of textual patterns in four subject categories

An in-depth analysis of four subject categories was conducted to characterize them according to the specific role of acknowledgments in each discipline. It is interesting that both funding acknowledgment and sub-authorship collaboration data contribute to define the specific pattern of each category. Our aim was to extract knowledge embedded within the text. Knowledge is expressed in words, so a lexicometric analysis

[29] In 2010, there were no English-written publications by Spanish authors in seven WoS subject categories.

establishing a statistical relationship among lexical units was carried out. The subject categories selected for these purposes were the following: cardiac and cardiovascular systems, economics, evolutionary biology, and statistics and probability (Table 9.2). They differ in their subject area, their theoretical versus experimental orientation, and the basic versus applied nature of the research, as measured through the research level indicator. These features were taken into account on the assumption that they might have an influence on the type of information to be included in the acknowledgment section of papers.

Table 9.2. Number of papers, presence of funding acknowledgments and average research level by subject category.

Subject category (WoS)	Subject Area	Broad area	No. Papers in WoS (2010)	% Papers with FA	Research level (mean)
Cardiac & Cardiovascular Systems	Clinical Medicine	Health Sciences	380	52.4%	1.9
Economics	Social Sciences	Social Sciences	546	12.8%	2.8
Evolutionary Biology	Agric., Biol. & Env.	Natural Sciences	271	88.5%	4
Statistics & Probability	Mathematics	Exact Sciences	294	75.5%	2.6

To analyze the text appearing in the acknowledgment section, textual data analysis was used (Lebart & Salem, 1994; Lebart et al., 2000) and the frequency of occurrence of words was obtained. The corpus was segmented into minimal units for frequency calculation. The processing of textual data and the building of a lexical table were carried out by means of Lexico 3 (Lamalle, Martínez, Fleury, & Salem, 2003). Because the software considers variant forms of a given term as different terms, a previous text normalization was performed to prevent such problems. Accordingly, orthographic variations were unified: for example, spelling variations (center vs. centre); variant forms resulting from slashes, hyphens (co-author vs. coauthor), or punctuation marks (WHO or W.H.O.); and variations resulting from the use of capital letters (capital letters were maintained only in personal and institutional names). Acronyms were also revised as a potential distorting element (for example, FIS or Fondo de Investigaciones Sanitarias). Stop words, that is, words with little semantic content that do not provide useful information for the analysis, were removed (articles, pronouns, prepositions, conjunctions, auxiliary and modal verbs). Midlevel "lemmatization" was applied (Bolasco, 1992), which means that the different inflected forms of a word were grouped to its lemma, allowing them to be analyzed as a single item: verbs into their infinitive form (e.g., supports, supported, supporting = support) and nouns into their

singular form. In addition, words with a different lemma but with equivalent semantic content were grouped together (for example, JAE, FPU, and FPI were grouped together under the *fellowship* entry since they are acronyms for different programs of fellowship grants in Spain). Finally, personal or institutional names and project numbers have not been taken into account for the analysis because our focus is on identifying acknowledgment patterns and making possible sub-authorship inferences. As a matter of fact, we are more interested in the reasons underlying the acknowledgments than in the identification of the individuals and institutions acknowledged. A threshold of 10 occurrences in the corpus was set in order to select the words to be included in the lexical table. Different frequency thresholds were tested, and the position of words and subject categories in the factorial planes remained quite stable. A lexical table is a cross-tabulation formed by the words and the four selected subject categories and the number of cooccurrences for each category, so that cell (i, j) contains the number of occurrences for word i in the FA of the category j. The matrix obtained has the form X_{93x4} (93 words and four subject categories).

Correspondence Analysis (CA) is a multivariate technique for displaying the rows and columns of a two-way contingency table as points in a low-dimensional space, so that the positions of the row and column points are consistent with their associations in the lexical table. This method was proposed by Benzécri (1973) and reviewed by Escofier and Pagès (1992) and Lebart and Salem (1994), among others. In our study, CA was applied to a data matrix of words and subject categories to find two vector spaces, one representing the words and the other the categories. For distances between two points corresponding to words and two points corresponding to subject categories to make sense, row-profile and column-profile tables are calculated to show the relative frequency distributions of a line of the table (row or column) regarding its total marginal.

With the profiles for the construction of the points, the differences between the distributions of acknowledgments in the text samples are measured by χ^2 distances (which are weighted Euclidean distances between normalized rows with weights inversely proportional to the square roots of the column totals) associated with the matrix into orthogonal factors. The proximity between subject categories represents similarities between them; that is, if two categories are very close in the projection, they are characterized by the same words. The distances between words and subject categories should be interpreted only in barycentric terms. In general, the points near the origin are underrepresented (Berthier & Bouroche, 1975). Note that the weight assigned to the lines of the matrix is inversely proportional to its total marginal. Therefore, words with the highest frequency rates are placed near the origin, whereas those with lower frequency rates will move away from the center of gravity of the axes. Because the WoS includes FA when papers include information about their

funding, words with the highest frequency ranking are expected to be related to funding, so lower frequency terms will be the ones relevant to explore different patterns in sub-authorship information.

For the interpretation of CA seeking maximum inertia plans, the number of factors required to represent the information properly must be addressed. The importance of each axis is measured through a percentage of inertia (i.e., variance) represented by an eigenvalue (λ), which measures the inertia of each of the principal axes, that is, $\lambda_1/\Sigma\lambda_n$ measures the inertia absorbed by the first axis and $(\lambda_1 + \lambda_2)/\Sigma\lambda_n$ measures the inertia absorbed by the planes 1–2. Total inertia is equal to the sum of all principal inertias $(\lambda_1, \lambda_2, \lambda_3 \ \ldots \ \lambda_n)$. In addition, several measures are important to obtain a correct interpretation. The relative contribution of a factor to the element is the relative variability of the variable (subject category) accounted by that factor. These relative contributions tell us how the information is distributed across the axes. In addition, for a point (row or column) on a factorial plane, the quality of representation can be defined by adding the two relative contributions of these factors to the element. Only words and subject categories with a high quality of representation can be properly interpreted. In this study, the quality of representation is rated on a scale ranging from 0 to 1,000 points. Words with a quality of representation below 400 points are not represented in factorial planes. In addition, Ward's hierarchical clustering method was applied using factor scores to identify acknowledgment patterns of subject categories based on similar lexical features. The statistical analysis was run in MultBiplot software (Vicente-Villardón, 2010). A flow diagram of the process is displayed in Figure 9.1.

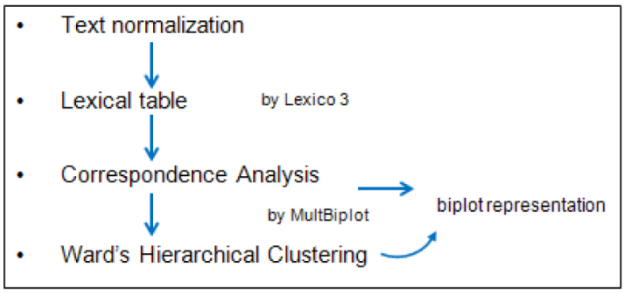


Figure 9.1. Stages in the textual analysis of acknowledgments.

9.4. Results

The scientific papers published in English by Spain-based researchers in 2010 total 43,360 publications, among which 27,774 (64%) present FA data. Our analyses are

based on citable items[30] (38,257), with FA being present in 72.6% of them (Table 9.3). Differences in FA presence by genre can be observed in Table 9.3. The results are presented in two distinct sections devoted to (a) the analysis of FA presence by subject area and (b) the analysis of textual patterns by subject category.

Table 9.3. FA presence by genre.

Genre	Without FA	With FA	Total
Articles	8,113 (23.4%)	26,577 (76.6%)	34,690
Reviews	642 (35%)	1,191(65%)	1,833
Proceedings papers	1,728 (99.7%)	6 (0.3%)	1,734
Total citable items	10,483 (27.4 %)	27,774 (72.6%)	38,257

Note. $\chi^2 = 4{,}636.6$, $p < .01$.

9.4.1 Analysis of FA presence by subject area

The presence of FA in English-language papers published by Spanish authors varies by subject area (Table 9.4.). It is low in the humanities (20%) and social sciences (28%) and intermediate in clinical medicine (59%), whereas it stands above 73% in the remaining areas. Physics and chemistry have the highest percentages of publications with FA (above 81%). Figure 9.2 shows the distribution of the category share of papers with FA by subject area, which is quite similar across experimental areas (e.g., physics, chemistry, mathematics) and presents a higher level of dispersion in engineering and technology (from 34% in transportation to 87% in mathematical and computational biology) and biomedicine (from 61% to 95%). The case of the multidisciplinary area should be analyzed with care, because only two subject categories make up the whole area. Clinical medicine shows a more dispersed pattern in FA frequency, which ranges from surgery (21%) to toxicology (83%), whereas social sciences presents a rightward-skewed distribution, from sociology and international relations (0%) to biological psychology (85%). Finally, humanities also shows a rightward-skewed distribution, from linguistics (8%) to archaeology (52%).

[30] The "proceedings paper" genre refers to articles which have been previously presented in a conference. The two genres ("article" and "proceedings paper") are assigned to these papers by *Thomson Reuters*.

Table 9.4. Presence of FA by subject area.

Subject area	Total journals			Q1 journals		
	No. Papers	No. Papers with FA	% Papers with FA	No. Papers	No. Papers with FA	% Papers with FA
Agric., Biol. & Envir.	7,974	6,473	81.1	4,907	4,172	85.0
Biomedicine	7,934	6,413	80.8	4,385	3,852	87.8
Chemistry	5,951	4,996	84.0	4,018	3,473	86.4
Clinical Medicine	7,609	4,525	59.4	3,969	2880	72.6
Engineer. & Technol.	8,573	6,298	73.5	5,212	4,183	80.3
Humanities	356	72	20.2	74	35	47.3
Mathematics	2,220	1,695	76.4	931	764	82.1
Multidisciplinary	655	520	79.4	456	419	91.9
Physics	7,064	5,779	81.8	4,327	3,821	88.3
Social Sciences	2,882	805	27.8	1,143	542	47.4
Total	38,257	27,774	72.6	21,058	17,209	81.7

Note. The sum of publications exceeds the actual total because there are journals assigned to more than one area.

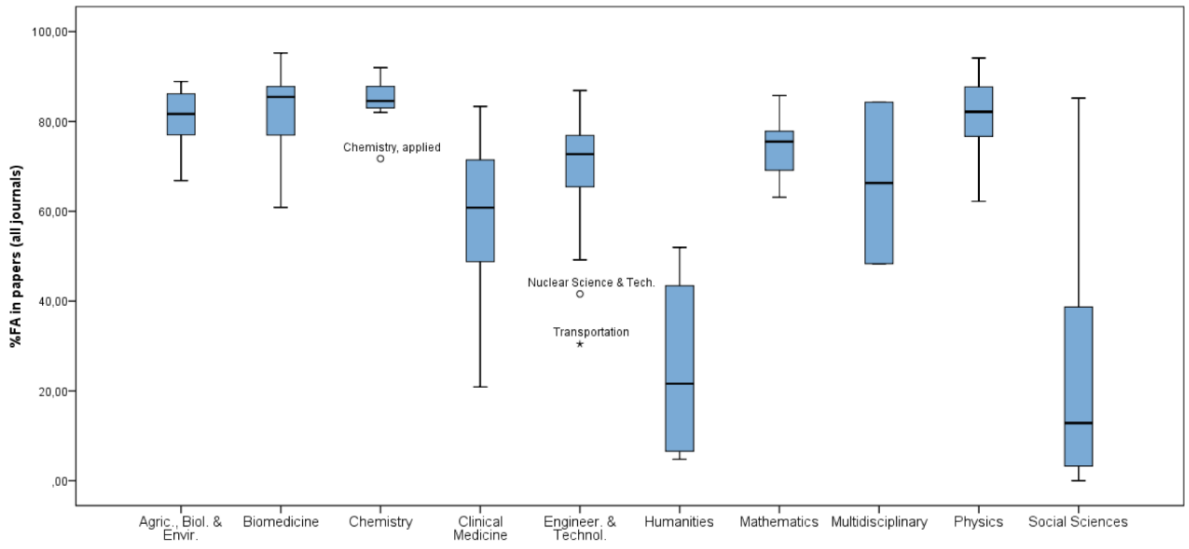


Figure 9.2. Distribution of the category share of papers with FA by subject area.

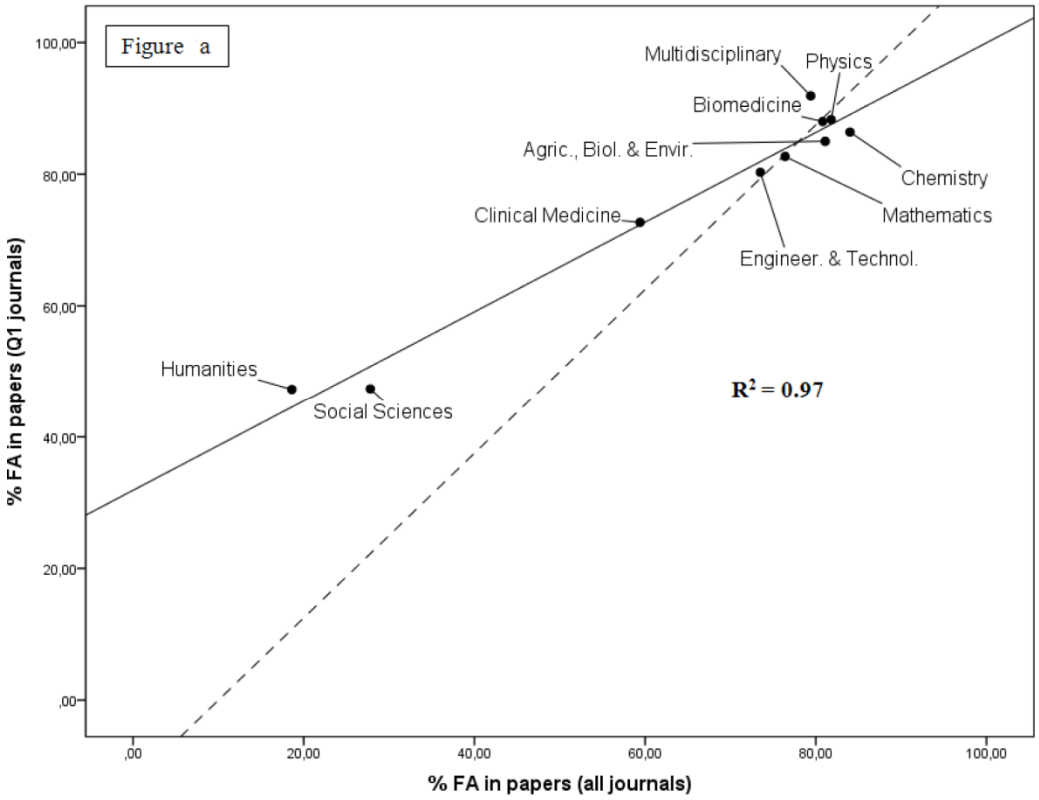
Note: Only subject categories with at least 30 papers are shown (*n* = 196).

9.4.2 Influence of journal prestige

Interestingly, the presence of FA in Q1 journals is higher than in the total set of journals (82% vs. 73%). This applies to all subject areas, but the differences are

especially significant in the humanities[31] (47% of FA in Q1 journals vs. 20% in total journals) and social sciences (47% vs. 28%; Table 9.4).

A high correlation is observed between the percentage of papers with FA in all journals and in the set of Q1 journals by subject area (Figure 9.3a). To explore the extent to which there are differences by subject category within a given area, the percentage of papers with FA by category is also shown (Figure 9.3b); colors are used to identify the subject categories in a given area). A strong and positive correlation is observed at both levels, subject areas, ρ = .82, $p < .01$, and subject categories, ρ = .91, $p < .01$. Most areas and categories are placed above the diagonal line in the graph (dashed line), which means that the percentage of FA in Q1 journals tends to be higher than in total journals.



[31] It should be noted that the impact factor is not calculated by *Thomson Reuters* for journals only included in Arts & Humanities. It is only available for some journals also included in the SSCI.

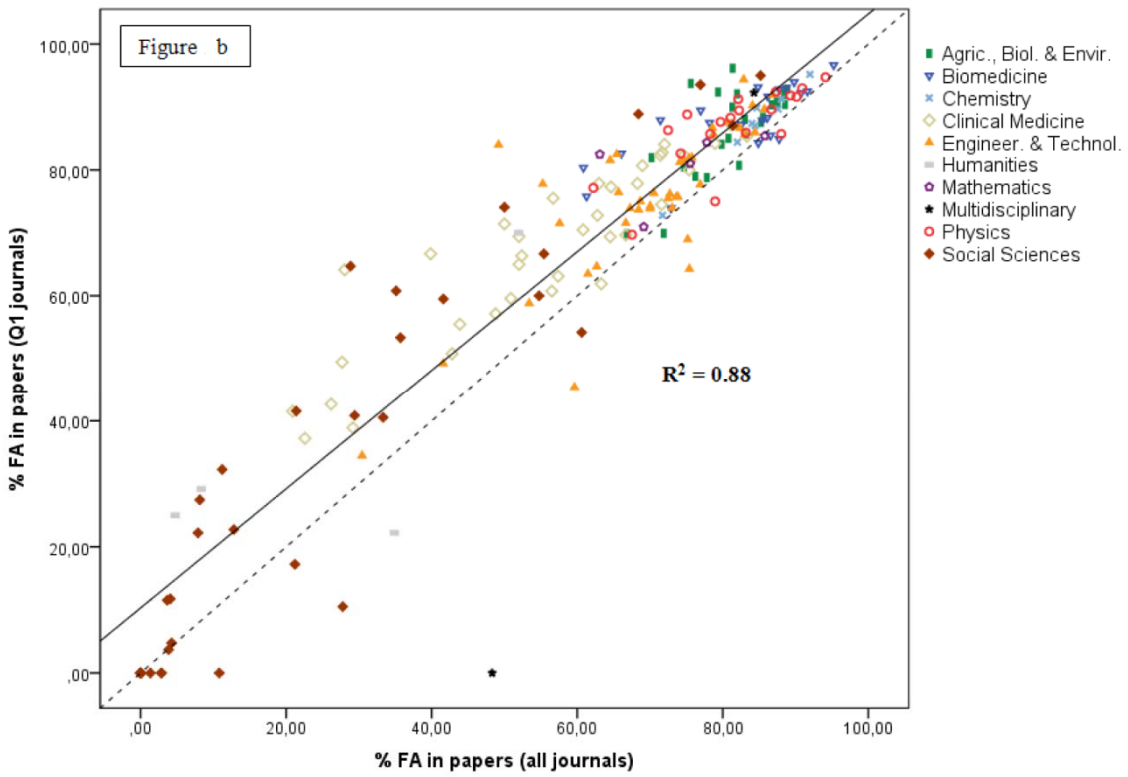


Figure 9.3. Relationship between the percentage of papers with FA in total journals and in Q1 journals by subject area (a) and subject category (b).

Notes: Only subject categories with at least 30 papers are shown in the scatter plot (n = 196). The solid line represents the regression line. The dashed line represents the diagonal in the graph.

We can see in Figure 9.3b that subject categories in a given subject area tend to group together. However, it is worth noting that the greatest scattering of categories is observed in social sciences and clinical medicine. A more detailed analysis of the social sciences area is shown in Figure 9.4, in which the percentage of papers with FA in the total set of journals and in the subset of Q1 journals by subject category is shown (excluding those without FA). As mentioned, FA is more likely to appear in Q1 journals, although the frequency varies largely by category. The areas closer to experimental sciences are more likely to include FA; see, for example, psychology, biological (95%), health care sciences and services (89%), and geography, physical (87%). Almost no FA presence is observed in other subject categories such as psychology, educational; business; and history of social sciences. Interdisciplinary differences with regard to the level of economic resources needed to conduct research can be a determining factor of FA presence.

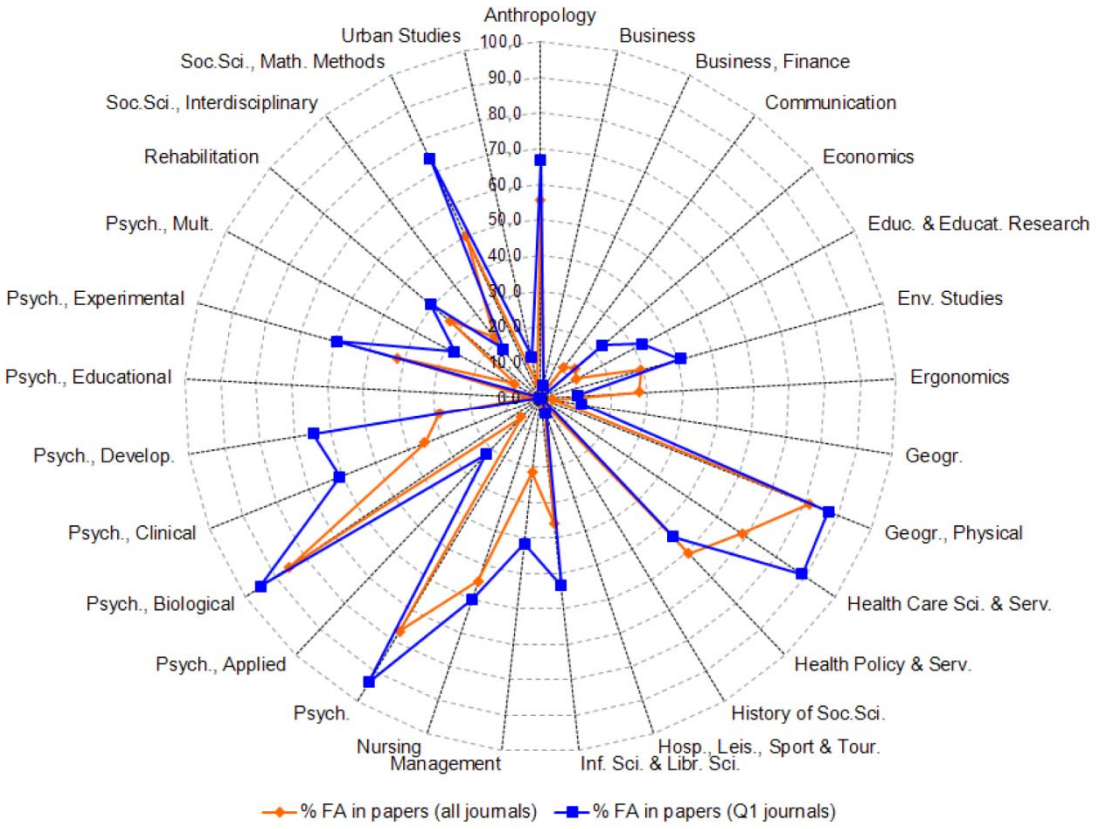


Figure 9.4. FA presence in social sciences subject categories: total journals and Q1 journals.

9.4.3. Influence of the number of authors

The relationship between the number of authors per paper and the presence of the acknowledgment section is analyzed. Assuming that team size grows with the complexity of research, the need for infrastructure, and the level of required economic support, FA presence is expected to increase with the number of authors per paper. Table 9.5 shows the average number of authors per paper by subject area depending on whether the FA section is present or not. The average number of authors tends to be higher for the set of papers including FA, the differences being statistically significant in engineering and technology, clinical medicine, social sciences, and humanities (Mann-Whitney's *U* test, $p < .01$).

Table 9.5. Average number of authors related to the presence of FA by subject area (mean ± Standard Deviation [*SD*]).

	No. authors/paper		
	M ± *SD*		
Subject area	Without FA	With FA	*p*-value
Agric., Biol. & Envir.	5.0 ± 1.6	5.0 ± 0.9	NS
Biomedicine	6.0 ± 1.3	6.8 ± 1.6	NS
Chemistry	5.0 ± 1.0	5.1 ± 0.4	NS
Clinical Medicine	6.1 ± 1.2	8.3 ± 2.2	< 0.01
Engineer. & Technol.	4.1 ± 1.3	5.2 ± 4.4	< 0.01
Humanities	1.6 ± 0.8	3.3 ± 1.4	< 0.01
Mathematics	2.8 ± 0.3	3.0 ± 0.3	NS
Multidisciplinary	6.4 ± 5.0	7.9 ± 6.1	NS
Physics	13.0 ± 19.4	14.5 ± 19.7	NS
Social Sciences	3.3 ± 1.2	5.1 ± 2.8	< 0.01

The relationship between the number of authors per paper and FA frequency varies by subject area (Figure 9.5). It is worth noting that the share of papers with FA increases almost linearly with the number of authors in some areas, such as clinical medicine, social sciences, and humanities. In the remaining areas, the highest increase in FA presence is observed from one-author to two-author papers, showing a very small increase thereafter. The multidisciplinary area shows a mixed pattern: a notable surge in FA presence is observed from one-author to two-author papers, but FA increases progressively with the number of authors involved.

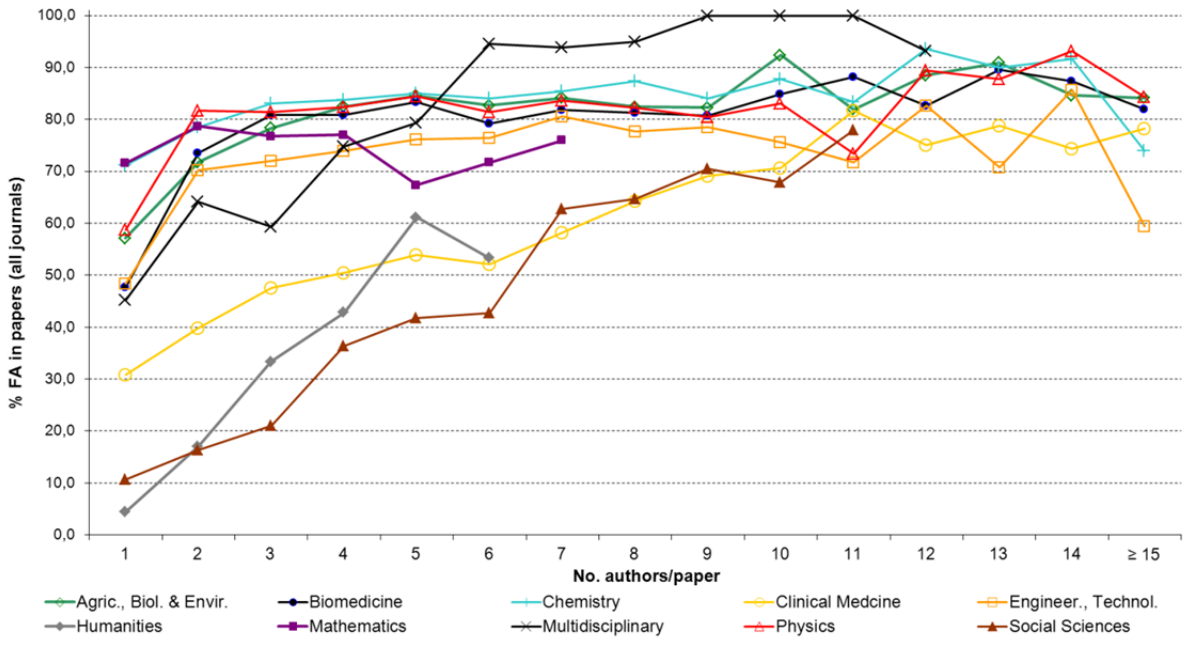


Figure 9.5. Presence of FA by number of authors per paper and subject area.

9.4.4. Influence of the research level

On average, papers with FA tend to show a slightly higher basic research level than the rest of the papers, although statistically significant differences are observed only for clinical medicine (Table 9.6).

Table 9.6. Average research level related to the presence of FA by subject area.

Subject area	**Research level**		
	M ± *SD*		*p*-value
	Without FA	With FA	
Agric., Biol. & Envir.	2.9 ± 0.8	2.9 ± 0.8	NS
Biomedicine	3.2 ± 0.4	3.3 ± 0.4	NS
Chemistry	3.2 ± 0.9	3.3 ± 0.9	NS
Clinical Medicine	1.8 ± 0.4	2.1 ± 0.5	< 0.01
Engineer. & Technol.	1.8 ± 0.6	2.0 ± 0.5	NS
Mathematics	2.6 ± 0.7	2.7 ± 0.7	NS
Multidisciplinary	2.5 ± 1.1	2.9 ± 1.3	NS
Physics	3.3 ± 0.5	3.4 ± 0.5	NS
Social Sciences	1.7 ± 0.8	2.1 ± 0.9	NS

Note: Humanities journals are not shown because the research level is only calculated for SCI and SSCI journals.

9.4.5. Analysis of textual patterns by subject category

A textual data analysis of the FA section was carried out for 1,491 papers published in the four selected subject categories. These categories differ dramatically in terms of FA presence: it ranges from 12.8% in economics to 88.5% in evolutionary biology (Table 9.2). In our study, the entire corpus comprises 50,710 word occurrences ("running words"), of which 10,936 are different forms ("types"; Table 9.7). Hapax legomena (those words with only one occurrence in the corpus) total 7,124 (14% of running words; 65% of types). It is important to note that, although the semantic richness of the acknowledgments is not very high because of the specific role of this section in the papers, the number of hapax legomena is high because of references to projects and persons.

Table 9.7. Lexical features of the corpus.

	Cardiac & Cardiovascular Systems	Economics	Evolutionary Biology	Statistics & Probability	**Corpus**
No. Running words	11,609	4,734	23,600	10,767	50,710
No. Types	2,436	1,351	5,104	2,045	10,936
Max. Word frequency	605	344	1,287	755	2,991
No. of hapax legomena	1,509	844	3,456	1,315	7,124

CA was applied to the lexical table that we obtained, which is a cross-tabulation including word occurrences for each of the four subject categories. The first two axes retain 92.9% of all the information contained in the lexical table.

Relative contributions of the factor to the element for the columns (Table 9.8) show that axis 1 is determined by the cardiac and cardiovascular systems subject category, whereas axis 2 is configured by the rest of categories, statistics and probability being its leading contributor. With regard to the rows, words related to economic issues have the highest contributions (*consultant*, *fee*, *employee*, *honorary*) in axis 1, whereas axis 2 is characterized by words that reflect some type of contribution (*analysis*, *collect*, *assistance*, *technical*).

Table 9.8. Relative contributions of the factor to the element for subject categories.

Subject categories	Axis 1	Axis 2	Axis 3
Statistics & Probability	47	**813**	139
Cardiac & Cardiovascular Systems	**992**	7	1
Economics	120	**537**	**343**
Evolutionary Biology	**374**	**626**	0

CA results shown in Figure 9.6 reveal differences in the lexical patterns of the subject categories selected. Economics and statistics and probability are found close to each other in the first quadrant of the spatial plot, suggesting that they present similar lexical profiles and similar acknowledgment patterns. Evolutionary biology stands in the fourth quadrant because it is characterized by different words. These subject categories are close in the projection to axis 2, which is characterized by words denoting some type of contribution, support, or process involving technical or research

work. Conversely, cardiac and cardiovascular systems is located in the third quadrant close to axis 1, where some words about grants and economics present a high level of contributions.

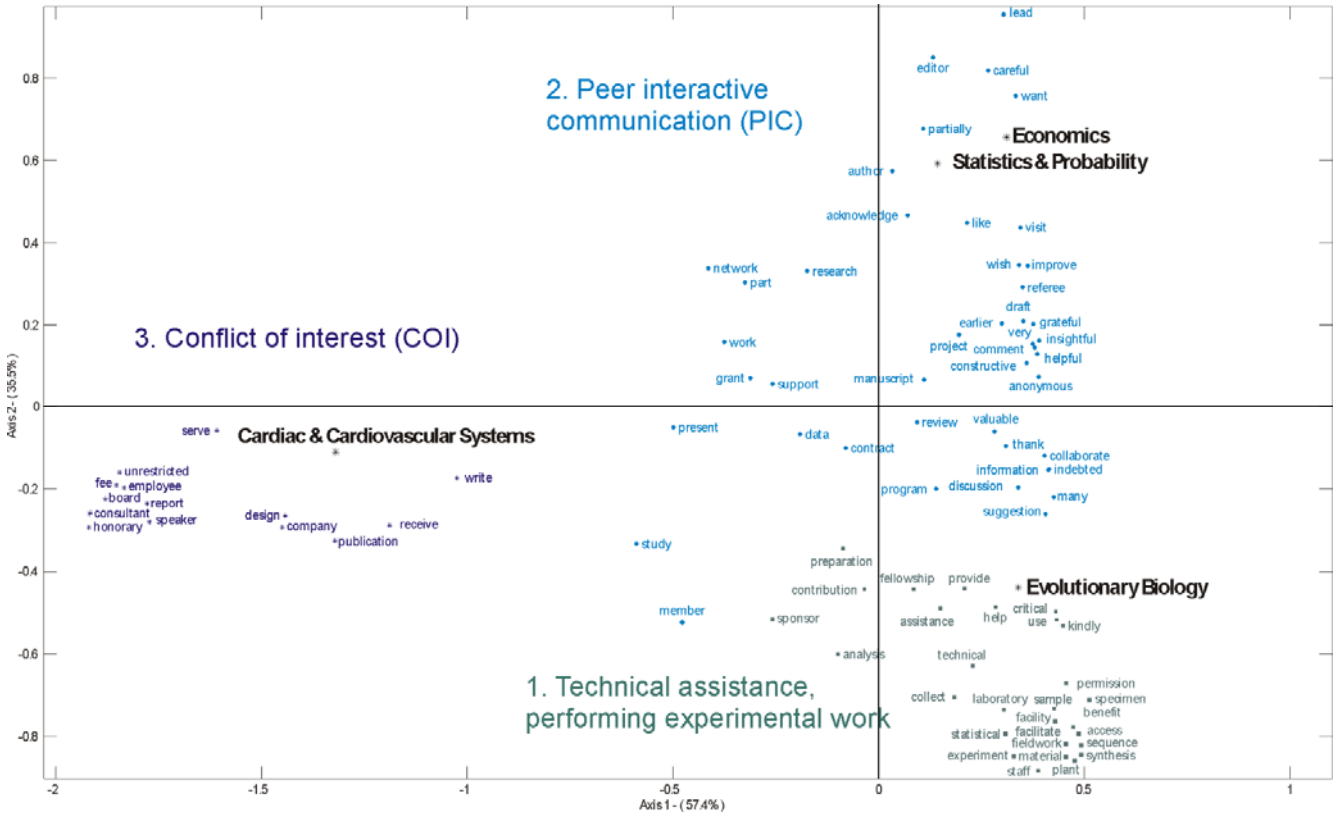


Figure 9.6. CA representation of the different clusters obtained on the principal factorial planes 1–2.

Ward's hierarchical clustering method using factor scores was applied to describe the pattern of the different subject categories, which includes both funding and sub-authorship collaboration data (Figure 9.6). Quality of representation is over 93% in planes 1–2 for clusters detected. Three clusters have been identified according to the lexical features of the corpus.

- ✓ Cluster 1 is formed by words that reveal gratitude for allowing access to facilities or assistance in sample collection, statistical analysis, or laboratory work. These contributions can be defined as technical assistance and performing experimental work. It is the pattern found in evolutionary biology.

- ✓ Cluster 2 recognizes some intellectual debt that contributes to improve the quality of research. It includes words such as *comment*, *improve*, *draft*, *insightful*, *helpful*, *careful*, *anonymous*, and *referee* and reflects cognitive, moral, and editorial support received by the authors, that is, the "peer interactive communication"

concept (PIC) introduced by McCain (1991). This is the pattern described for economics; and statistics and probability.

- ✓ Cluster 3 is formed by words that reveal an existing concern about potential conflicts of interest in the conduct and publication of research. It includes words such as *honorary*, *fees*, and *consultants*, in relation to the collaboration between academic scientists and industry. This is the pattern described in cardiac and cardiovascular systems, which is a clinical discipline that differs completely from the rest of categories analyzed in the nature of the information recorded in the acknowledgment section of papers.

9.5.Discussion

Our study analyzes acknowledgments in the English-language scientific publications of Spain-based researchers in WoS in 2010 and confirms the existence of differences in FA presence by subject area, by genre, by journal prestige, and in a specific area, by the basic versus applied nature of research. Moreover, text mining and multivariate analysis have proved useful in discovering interfield differences in the patterns of FA based on similar lexical profiles. Several disciplinary features, such as their theoretical versus experimental nature, the need for economic support, the relevance of teamwork, and the dependence on complex facilities contribute to build differentiated patterns of acknowledgments.

An FA section was present in approximately 73% of the publications written in English by Spain-based researchers in 2010, which is a higher rate than that found for Spanish publications in previous studies (Costas & van Leeuwen, 2012; Wang et al., 2012). Several methodological factors, such as the language and genre of the studied papers, may contribute to explain these differences. On the one hand, the fact that only papers written in English are taken into account in our study may contribute to explain the higher values of FA presence observed, because weaker funding acknowledgment rates have been found for papers written in other languages in the literature (Salager-Meyer et al., 2009). On the other hand, our study focuses on articles and reviews, whereas studies including other genres such as editorials or letters, which very rarely contain acknowledgments, are likely to reveal a lower presence of FA (Costas & van Leeuwen, 2012).

Moreover, an additional factor should be borne in mind. When trying to explore potential differences in FA presence between papers written in English and in other languages by Spain-based researchers, we observed that funding acknowledgments were not recorded at all by the WoS for the set of papers not written in English. After submitting a query to *Thomson Reuters*, we verified that acknowledgments must be in

English to be captured and processed by the WoS. This observation has important implications; papers written in other languages could be considered as nonfunded, although the fact is that we have no information on this issue. Accordingly, the rate of nonfunded papers could be overestimated in those studies in which all papers (not only papers written in English) are considered. This can be especially significant in the more locally-oriented areas, such as social sciences or clinical medicine, in which the share of papers in languages other than English is higher.

In any case, it is interesting to point out that FA presence in Spanish publications was above the world average in 2009 in the study of Costas and van Leeuwen (2012), in which Spain held fifth position (after China, Sweden, South Korea, and Finland) in the world ranking of top FA frequency countries. The existence of mandates for explicit mention of funding agencies has been argued to explain a high share of acknowledged papers in some countries (Costas & van Leeuwen, 2012), Spain being one of them (BOE, 2008). The fact that FA is more likely to appear in articles than in reviews was also observed by Costas and van Leeuwen (2012) and Salager-Meyer et al. (2011). It is clear that articles including original research require more collaboration and infrastructure than reviews, which usually consist of elaboration of previous research on a specific topic. On the other hand, the low presence of FA in proceedings papers may be due to the fact that they tend to be shorter and to contain less detailed information than articles (González-Albo & Bordons, 2011).

9.5.1. Funding acknowledgment by subject area

Our study reveals important differences in FA presence by area. The lowest values of FA are found in humanities and social sciences and the highest in experimental fields, such as chemistry and physics, these results being consistent with previous studies (Cronin et al., 2003, 2004; Costas & van Leeuwen, 2012). The theoretical versus experimental nature of research and its technical complexity, which may require sophisticated infrastructures and teamwork, are joint factors determining the field's dependence on economic resources and, therefore, FA presence too. The low share of FA in the humanities and social sciences could also be influenced by cultural factors; the inclusion of formal acknowledgments in papers is a more widely established tradition in experimental fields (Costas & van Leeuwen, 2012).

The higher presence of FA described for the more basic fields in a previous study (Costas & van Leeuwen, 2012) is supported by our results, because engineering and technology and clinical medicine are the areas with the highest applied research level and the lowest FA rate (apart from social sciences). We do not know the extent to which this is due to the greater dependence of basic research on extramural funding or to the different sources of funding used by basic and clinical research. In any case, within any given area, significant differences in the research level of funded versus

nonfunded papers were observed only in the case of clinical medicine, in which funded papers present a more basic level, maybe because research more clearly oriented to clinical practice is more likely to be developed with intramural resources that are not specifically acknowledged.

9.5.2. Funding acknowledgment by journal prestige

An interesting finding from our study is the higher presence of FA in high-impact-factor journals (Q1), which suggests the higher quality of funded research. Among the possible underlying reasons for this, we can mention the more stringent peer review process applied to funded research or the fact that funding allows scientists to allocate more time to research, access better technology, or collaborate with more qualified scientists (Zhao, 2010). Publications with FA were also published in higher impact factor journals in the study of Costas and van Leeuwen (2012), in which they obtained a higher citation rate than nonfunded publications. A higher citation rate for publications with FA has also been described elsewhere (Lewison & Dawson, 1998; Zhao, 2010; Levitt, 2011). Although a positive relationship between the number of funding sources and the citation impact of papers has been suggested in the literature (Lewison & Dawson, 1998; Rigby, 2011), Rigby concludes in a recent study that the effect of the number of funding acknowledgments is weak and that it should not be considered a reliable indicator of research impact (Rigby, 2013). On the other hand, Zhao (2010) advocates a cautious approach, because some of the most cited papers include no funding acknowledgments (Zhao, 2010). Unfortunately, citation data were not analyzed in our study and, therefore, we cannot provide any new evidence on this issue.

In this study, the biggest difference in FA presence between all journals and Q1 journals was observed in the humanities, social sciences, and clinical medicine, in that order. These areas show the lowest share of papers with FA, but FA presence reveals its greatest increase if only high-impact-factor journals are considered (Q1 journals). We could argue that research in these areas is more locally oriented (Archambault, Vignola-Gagne, Cote, Lariviére, & Gingras, 2006; González-Alcaide, Valderrama-Zurián, & Aleixandre-Benavent, 2012), and perhaps funding is more frequently associated with more international research topics within each area, which are more easily placed in top-rank journals. However, a deeper analysis of the data would be required to confirm this hypothesis.

Because national journals are rarely placed in the first quartile of the impact factor journal ranking within their disciplines, a high volume of publications in them might contribute to the lower FA presence in the total set of papers compared with the Q1 set. However, the share of papers in Spanish journals is very low in this study, because only English-language papers were considered (it accounts for 2% of papers; ranging

from 0.1% of papers in chemistry to 7% in social sciences and 12% in humanities). In any case, the study of potential differences between national and international journals in FA presence remains an interesting topic for further research. A lower presence of FA in non-English journals has been described in the literature (Salager-Meyer et al., 2009), where it was attributed to a less strict commitment to comply with international authorship guidelines. The extent to which this applies to journals published in Spain and whether there are differences according to their language remain open questions. However, for the time being, they cannot be addressed through the WoS database because acknowledgment data are not recorded for Spanish-language journals.

9.5.3. Funding acknowledgment by number of authors

In our study, higher numbers of authors are observed for papers with FA compared with those without FA. Different interrelated factors may contribute to explain this tendency. On the one hand, this may be due to the fact that more complex research requiring more infrastructure and teamwork is more likely to be funded because it cannot be conducted without economic support. On the other hand, it should be mentioned that scientists in Spain are encouraged to form teams when they apply for research grants from the most important agencies as a means to foster team consolidation in the country, so that larger teams may stand a better chance in their quest for funding. The fact that collaborative research may be favored by funding agencies has been previously suggested in the literature (e.g., Zhao, 2010). Finally, it is interesting to point out that the higher number of authors per paper of funded research could be an influential factor in the final quality of the papers and help explain why they are more often found in Q1 journals, given that a positive relationship between the number of authors and the impact of research has been formerly described in the literature (Gazni & Didegah, 2011; Bordons, Aparicio, & Costas, 2013).

It should also be noted that in most subject areas the highest increase in FA presence is observed when we move from single-authored to multiauthored papers, because the former are less likely to receive funding, and because such research is thought to be less dependent on infrastructure. The situation is somewhat different in the areas of the social sciences, humanities, and clinical medicine, which show the lowest overall FA presence, although it tends to increase linearly with the number of authors. In the first two areas, single-authored papers represent an important share of research, and funding may have a critical role in boosting collaboration and multiauthored papers. In the case of clinical medicine, many papers might have been written by physicians as a result of their professional practice at medical sites with no special extramural funding, and yet as the number of authors increases so does the probability of having

extramural funding, especially for larger teams, who may be involved in clinical studies frequently supported by commercial companies.

9.5.4. Acknowledgment patterns by subject category

Despite the fact that only four categories were studied, the existence of interfield differences in the textual pattern of the acknowledgments is confirmed here, and different roles played by acknowledgments depending on the field have been identified. These differences are dependent mainly on the type of research, but other influences, at both the local and the global levels, such as social and cultural factors, also play a role.

Our study shows that PIC is a characteristic feature of the more theoretical or socially-oriented fields (i.e., statistics and probability, economics), whereas the recognition of technical aid (data collection and analysis) is more common in experimental research (i.e., evolutionary biology), and the mention of potential conflicts of interest occurs especially in the clinical field (i.e., cardiac and cardiovascular systems). With regard to "sub-authorship information," PIC acknowledgments are particularly relevant, in that they imply an intellectual debt and, as has been pointed out by other authors (McCain, 1991; Davis & Cronin, 1993), suggest an extra peer-review process before publication, which may enhance the quality of the final paper (Costas & van Leeuwen, 2012). The recognition of technical work is also important; it can prove essential for the development of experimental research in specific fields. Finally, the acknowledgment pattern described for the clinical field is not so clearly related to sub-authorship information; it deals mainly with personal or financial relationships (i.e., with commercial firms) that might potentially bias the authors' research and compromise the credibility of their publications.

Conflicts of interest related to project support are becoming ever more common in clinical medicine; scientists may receive funding from commercial firms and private foundations, which may interfere with their ability to analyze data independently, prepare, and publish the results. As a consequence, leading journals request authors to disclose the potential financial interests of sponsors and to describe their involvement in the research project, where appropriate. Although conflicts of interest may lead to biased research publications, a close relationship between the academic and the industrial sectors is beneficial for research (Stossel, 2012), and the disclosure of information on potential conflicts of interest contributes to strengthen article credibility and public confidence in research.

What emerges from our study is that the contents of the acknowledgment section vary largely by subject category and can contain very heterogeneous data. As the research process gains complexity (increasing role of teams and network-based research,

diversity of funding sources, more sophisticated administrative and legal frameworks, growing concern about ethical issues), so does the amount and variety of the information included in the acknowledgment section. Separating funding data, conflict of interest statements, and other types of acknowledgments (e.g., sub-authorship) is becoming the norm in some journals (*Lancet*, 2011; *Nature*, 2012) and may facilitate the task of authors when submitting papers, the flow of information to readers, and its study by interested scientists.

9.5.5. Authors, subauthors and contributors

The textual analysis of the FA section reveals acknowledgment patterns by discipline that are determined by both funding information and sub-authorship collaboration in science. Although our main interest was to identify acknowledgment patterns rather than specific collaborators, it is clear that subauthors providing technical and/or intellectual assistance to the research lie behind these patterns.

At this stage of the discussion, there emerges the interesting issue of trying to sort out the extent to which there is a clear delimitation between the kinds of contributions that deserve to be listed as co-authorship and those falling in the sub-authorship category. Different guidelines, most of them from journal editors, describe the qualifying criteria for authorship, but these are not universally accepted (Claxton, 2005). According to the guidelines of the International Committee of Medical Journal Editors (ICMJE, 2013), to be mentioned as an author, a scientist should not only make substantial contributions to the conception and design of the study or to the acquisition, analysis, and interpretation of data, but also participate in drafting the article or revising it critically as well as in the approval of its final version. Collaborators who do not fully comply with authorship criteria should be acknowledged (Claxton, 2005), but there are many signs that researchers are scarcely familiar with authorship criteria (Marusic, Bosnjak, & Jeroncic, 2011), which, in addition, may vary from one discipline, institution, or team to another. Although most author guidelines tend to privilege the creative and intellectual aspects of research over technical contributions, the interest of the latter is being increasingly recognized by scientists themselves (Winston, 1985; Hunt, 1991; Vinkler, 1993; Digiusto, 1994) and by journal editors (Wager, 2009). Indeed, the key issue is not only to decide what type of contribution (PIC, technical, etc.) deserves authorship but also what threshold of involvement is required.

The need to clarify the specific role of every author in a given piece of research has led some journals (for example, *Nature*[32]) to include a list of contributors, instead of authors, in their papers, where the contribution of each author to the research is

[32] http://www.nature.com/nature/authors/gta/

explicitly stated, thus blurring the differences between “authors” and “subauthors.” The contributorship system would open up new prospects for research in the field of collaboration issues in science but, although the ICMJE encourages journals to include contributor lists in their papers, only 10% of the biomedical journals had adopted the system by 2009 (Wager, 2009). In the meantime, the study of the acknowledgment section is an interesting option for an in-depth analysis of collaborative research practices, assuming that a sizeable part of them remains beyond the scope of the classical bibliometric indicators used to measure research collaboration, because these are based mainly on co-authorship analyses (Laudel, 2002) and are inadequate for the provision of a full and thorough image of collaboration in science.

9.5.6. Limitations of the study

In this study, the analysis of sub-authorship patterns is restricted to papers that present funding information, because only in those cases is the acknowledgment section of papers captured by the WoS. This is a limitation derived from the indexing policy of the WoS, but funding support is acknowledged in a large number of papers included in our study (73%), so we consider that the analysis of discipline patterns may yield reliable results.

The presence of acknowledgments in papers may be influenced by different factors such as the fact that it is not always mandatory in scientific journals (differences by field do exist), that all funding bodies may not be equally interested in having their support disclosed (commercial organizations may be reluctant to be mentioned for strategic reasons; Rigby, 2011), and that authors may differ in their propensity to acknowledge their influences and the assistance received. Although these limitations should be kept in mind, we consider that this study provides interesting insights into the presence and role played by acknowledgments in the English-language scientific output of a non-Anglophone country with a particular focus on interfield differences and sub-authorship information.

9.6. Conclusions and future research

Our study confirms the existence of differences in FA presence by subject area, genre, journal prestige, co-authorship, and, in a specific area, the basic versus applied type of research. Moreover, interfield differences in the nature of acknowledgments that go beyond financial support and include sub-authorship information were detected in the four subject categories under analysis. Extending the study to other research fields would allow us to categorize fields according to their acknowledgment patterns. Physics remains an attractive field for future research along with, in particular, large-scale science disciplines because of the specifics with regard to the contribution of

infrastructure to the development of large collaborative experiments that may be acknowledged in papers.

Moreover, this study opens up new avenues for future research. A comparative study of the presence of funded research in papers written in English and papers written in other languages is an interesting topic for further analysis, which has been addressed only at the level of specific journals (Salager-Meyer et al., 2009). On the other hand, multivariate analysis could be useful to delve into the relationships between acknowledgment presence and different features of research, such as collaboration, research level, and impact of research. The inclusion of citations received by papers along with the impact of publication journals would be desirable.

Certain developments in the way in which acknowledgment information is included in the WoS may enhance future research on the topic. First and foremost, the collection of funding acknowledgment data for all journals, regardless of their language, would be desirable. Second, the inclusion of the acknowledgment section in all WoS records, and not only when funding is acknowledged, would allow more global, comprehensive, and accurate studies. Finally, a better structuring of the acknowledgment information, including, for example, preestablished field subsections (in both journals and databases, as previously suggested by other authors, e.g., Cronin & Weaver, 1995), would allow us to discriminate between different types of information (i.e., financial data, conflict of interest disclosure, sub-authorship information) and facilitate automatic data processing. Today, there is a growing interest in the study of the acknowledgment section of papers because, as the research process grows in complexity, so does the amount and type of information included in the acknowledgments, revealing some of the particularities of research in each discipline.

Acknowledgments

This research was supported by the Spanish Ministry of Science and Innovation (CSO2008-06310) and the Spanish National Research Council (JAE predoctoral grant). We are very thankful to Javier Aparicio and Ignacio Santabárbara for their valuable assistance in data management as well as to Isabel Gómez and Purificación Galindo for their comments on a draft of this article. We especially thank two anonymous referees for their useful suggestions, which contributed significantly to improving the original manuscript.

References

Archambault, E., Vignola-Gagne, E., Cote, G., Lariviere, V., & Gingras, Y. (2006). Benchmarking scientific output in the social sciences and humanities: The limits of existing databases. *Scientometrics*, *68*(3), 329-342.

Benzécri, J.P. (1973). *L'Analyse de Données. Vol. 2. L'analyse des correspondances*. Paris: Dunod.

Berthier, F., & Bouroche, J.R. (1975). *Analyse des données multidimensionnelles*. Paris: Presses Universitaires de France.

Boletin Oficial del Estado (2008). Orden Pre/621/2008, por la que se regulan las bases, el régimen de ayudas y la gestión de la línea instrumental de actuación de proyectos de I + D + i, en el marco del Plan Nacional de Investigación Científica, Desarrollo e Innovación Tecnológica, 2008-2011. BOE num.59, 8 March 2008, pp.14, 217. <http://www.boe.es/boe/dias/2008/03/08/pdfs/A14205-14219.pdf> Accessed: 22/03/13.

Bolasco S. (1992). Sur différentes stratégies dans une analyse des formes textuelles: Une expérimentation à partir de données d'enquête, *Jornades Internacionals d'Analisi de Dades Textuals* (pp. 69-88). Barcelona: UPC.

Bordons, M., Aparicio, J., & Costas, R. (2012). Heterogeneity of collaboration and its relationship with research impact in a biomedical field. *Scientometrics*, doi: 10.1007/s11192-012-0890-7

Bordons, M., & Gómez, I. (2000). Collaboration networks in science. In B. Cronin & H.B. Atkins, (Eds.), The web of knowledge: A Festschrift in honor of Eugene Garfield (pp. 197-213). Medford, NJ: Information Today, Inc. & ASIS.

Costas, R., & van Leeuwen, T.N. (2012). Approaching the "reward triangle": general analysis of the presence of funding acknowledgments and "peer interactive communication" in scientific publications. *Journal of the American Society for Information Science and Technology*, *63*(8), 1647-1661.

Claxton, L.D. (2005). Scientific authorship part 2. History, recurring issues, practices, and guidelines. *Mutation Research*, *589*, 31-45.

Cronin, B. (1991). Let credits roll: a preliminary examination of the role played by mentors and trusted assessors in disciplinary formation. *Journal of Documentation*, *47(3),* 227-239.

Cronin, B. (1995) *The scholar's courtesy: The role of acknowledgments in the primary communication process*. Los Angeles: Taylor Graham.

Cronin, B., McKenzie, G., & Rubio, L. (1993). The norms of acknowledgment in four humanities and social sciences disciplines. *Journal of Documentation*, *49*(1), 29-43.

Cronin, B., McKenzie, G., & Stiffer, M. (1992). Patterns of acknowledgment. *Journal of Documentation*, *48*(2)*,* 107-122.

Cronin, B., McKenzie, G., Rubio, L., & Weaver-Wozniak, S. (1993). Accounting for influence: acknowledgments in Contemporary Sociology. *Journal of the American Society for Information Science*, *44*(7), 406-412.

Cronin, B., Shaw, D., & La Barre, K. (2003). A cast of thousands: co-authorship and sub-authorship collaboration in the 20th century as manifested in the scholarly literature of psychology and philosophy. *Journal of the American Society for Information Science and Technology, 54*(9), 855-871.

Cronin, B., Shaw, D., & La Barre, K. (2004). Visible, less visible, and invisible work: patterns of collaboration in 20th century Chemistry. *Journal of the American Society for Information Science and Technology, 55*(2), 160-168.

Cronin, B., & Weaver, S. (1995). The praxis of acknowledgment: from bibliometrics to influmetrics. *Revista Española de Documentación Científica*, *18*(2), 172-177.

Davis, C.H., & Cronin, B. (1993). Acknowledgments and intellectual indebtedness: a bibliometric conjecture. *Journal of the American Society for Information Science and Technology*, *44*(10), 590-592.

Digiusto, E. (1994). Equity in authorship: a strategy for assigning credit when publishing. *Social Science & Medicine*, *38*, 55-58.

Edge, D. (1979). Quantitative measures of communications in science: a critical review. *History of Science*, *17*, 102-134.

Escofier, B., & Pagès, J. (1992). *Análisis factoriales simples y múltiples. Objetivos, métodos e interpretación*. Bilbao, Spain: University of Basque Country.

Gazni, A., & Didegah, F. (2011). Investigating different types of research collaboration and citation impact: a case study of Harvard University's publications. *Scientometrics, 87*(2) 251-265. doi: 10.1007/s11192-011-0343-8

Giles, C.L., & Council, I.G. (2004). Who gets acknowledged: measuring scientific contributions through automatic acknowledgment indexing. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(51), 17599-17604.

Gillet, R. (1991). Pitfalls in Assessing Research Performance by Grant Income. *Scientometrics*, *22*, 253-263.

González-Albo, B., Moreno, L., Morillo, F., & Bordons, M. (2012). Bibliometric indicators for the analysis of the research performance of a multidisciplinary institution: the CSIC. *Revista Española de Documentación Científica*, *35*(1), 9-37. doi: 0.3989/redc.2012.1.851

González-Alcaide, G., Valderrama-Zurián, J.C., & Aleixandre-Benavent, R. (2012). The impact factor in non-English speaking countries. *Scientometrics, 92*(2), 297-311. doi: 10.1007/s11192-012-0692-y

Heffner, A.G. (1981). Funded research, multiple authorship, and subauthorship collaboration in four disciplines. *Scientometrics*, *3*(1), 5-12.

Hunt, R. (1991) Trying an authorship index. *Nature*, *352*(18), 187.

ICMJE. (2013). Uniform Requirements for Manuscripts Submitted to Biomedical Journals: Ethical Considerations in the Conduct and Reporting of Research: Authorship and Contributorship. <http://www.icmje.org/ethical_1author.html> Accessed: 04/02/13.

Kassirer, J.P., & Angell, M.A. (1991). On authorship and acknowledgments. *The New England Journal of Medicine*, *325*(21), 1510-1521.

Lamalle, C., Martínez, W., Fleury, S., & Salem, A. (2003). Lexico 3. Université de la Sorbonne nouvelle, Paris. <http://www.tal.univ-paris3.fr/lexico/>.

Lancet (2011). Information for authors. <http://download.thelancet.com/flatcontentassets/authors/lancet-information-for-authors.pdf>. Accessed: 01/03/13.

Laudel, G. (2002). Collaboration and reward. What do we measure by co-authorships? *Research Evaluation*, *11*(1), 3-15.

Lebart, L., & Salem, A. (1994). *Statistique textuelle*. Paris: Dunod.

Lebart, L., & Salem, A., Bécue, M. (2000). *Análisis estadístico de textos*. Lleida (España): Milenio.

Levitt, J.M. (2011). Are funded articles more highly cited than unfunded articles? A preliminary investigation. *Proceedings of ISSI 2011: The 13th Conference of the International Society for Scientometrics and Informetrics* (pp.1013-1015). Durban: ISSI.

Lewison, G., & Dawson, G. (1998). The effect of funding on the outputs of biomedical research. *Scientometrics*, *41*(1-2), 17-27.

Lewison, G., & Markusova, V. (2010). The evaluation of Russian cancer research. *Research Evaluation 19*(2), 129-144.

Marusic, A., Bosnjak, L., & Jeroncic, A. (2011). A systematic review of research on the meaning, ethics and practices of authorship across scholarly disciplines. *Plos One*, *6*(9), 1-17. doi: 10.1371/journal.pone.0023477

McCain, K.W. (1991). Communication, competition, and secrecy: the production and dissemination of research-related information in genetics. *Sciences, Technology & Human Values*, *16*(4), 491-516.

Melin, G., & Persson, O. (1996). Studying research collaboration using co-authorships. *Scientometrics 36*(3), 363-377.

Narin, F., Pinski, G., & Gee, H.H. (1976). Structure of the biomedical literature. *JASIS, 27*(1), 25-44.

Nature (2012). Guide to publication policies of the Nature journals. <http://www.nature.com/authors/gta.pdf>. Accessed: 01/03/13.

Noma, E. (1986). Subject Classification and influence weights for 3000 journals. Research report under CHI and NIH contracts. New Jersey: Computer Horizons Inc. Research.

Patel, N. (1973). Collaboration in the professional growth of American Sociology. *Social Science Information*, *12*(6), 77-92.

Rigby, J. (2011). Systematic Grant and funding body acknowledgment data for Publications: new dimensions and new controversies for research policy and evaluation. *Research Evaluation*, *20*(5), 365-375. doi: 10.3152/095820211X13164 38967039

Rigby, J. (2013). Looking for the impact of peer review: does count of funding acknowledgments really predict research impact? *Scientometrics*, *94*, 57-73. doi: 10.1007/s11192-012-0779-5

Salager-Meyer, F., Alcaraz-Ariza, M.A., Luzardo-Briceño, M., & Jabbour, G. (2011). Scholarly gratitude in five geographical contexts: a diachronic and cross-generic approach of the acknowledgment paratext in medical discourse (1950–2010). *Scientometrics*, *86*, 763–784. doi: 10.1007/s11192-010-0329-y

Salager-Meyer, F., Alcaraz-Ariza, M.A., & Pabón-Bervesí, M. (2009). "Backstage solidarity" in Spanish and English-written medical research papers. Publication context and the acknowledgment paratext. *Journal of the American Society for Information Science and Technology*, *60*(2), 307-317.

Stossel, T.P. (2012). What's wrong with COI? *The Scientist.* <http://the-scientist.com/2012/06/12/opinion-whats-wrong-with-coi/> Accessed: 20/12/12.

Thomson. <http://wokinfo.com/products_tools/multidisciplinary/webofscience/fundingsearch/>

Vicente-Villardón, J.L. (2010). MultBiplot: A package for Multivariate Analysis using Biplots. Departamento de Estadística. Universidad de Salamanca. <http://biplot.usal.es/ClassicalBiplot/index.html>.

Vinkler, P. (1993). Research contribution, authorship and team cooperativeness. *Scientometrics*, *26*(1), 213-230.

Wager, E. (2009). Recognition, reward and responsibility: Why the authorship of scientific papers matters. *Maturitas*, *62*(2), 109-112.

Wang, X., Liu, D., Ding, K., & Wang, X. (2012). Science funding and research output: a study on 10 countries. *Scientometrics*, *91*(2), 591-599. doi: 10.1007/s11192-011-0576-6

Wang, J., & Shapira, P. (2011). Funding acknowledgment analysis: an enhanced tool to investigate research sponsorship impacts: the case of nanotechnology. *Scientometrics*, *87*, 563-586. doi: 10.1007/s11192-011-0362-5

Winston, R.B. (1985). A suggested procedure for determining order of authorship in research publications. *Journal of Counseling and Development*, *63*(8), 55-518.

Zhao, D.Z. (2010). Characteristics and impact of grant-funded research: a case study of the library and information science field. *Scientometrics, 84*, 293-306.

ANEXOS

Anexo I. Áreas y disciplinas *Web of Science*.

Agricultura, Biología y Medio Ambiente
Agricultura y Ganadería
Agricultura, Multidisciplinar
Agronomía
Biodiversidad
Biología
Biología de la Evolución
Biología Marina y de Aguas Continentales
Biotecnología y Microbiología Aplicada
Botánica
Ciencia del Suelo
Ciencia y Tecnología de los Alimentos
Ecología
Entomología
Horticultura
Ingeniería Agrícola
Limnología
Medio Ambiente
Micología
Ornitología
Pesca
Política y Economía Agrícola
Recursos Hídricos
Silvicultura
Veterinaria
Zoología

Ciencias Sociales
Administración de Empresas
Administración Publica
Antropología
Biblioteconomía y Documentación
Ciencias Políticas
Ciencias Sociales Interdisciplinares
Ciencias Sociales y Biomedicina
Ciencias Sociales, Métodos Matemáticos
Comunicación
Criminología y Ciencia Penal
Demografía
Derecho
Economía
Economía Financiera
Economía, Negocios
Educación e Investigación Educativa
Educación Especial
Enfermería

Biomedicina
Anatomía y Morfología
Biofísica
Biología Celular
Biología del Desarrollo
Biométodos
Bioquímica y Biología Molecular
Ciencias del Comportamiento
Endocrinología y Metabolismo
Farmacología y Farmacia
Fisiología
Genética y Herencia
Ingeniería Tisular y Celular
Inmunología
Medicina, Investigación
Microbiología
Microscopía
Neurociencias
Parasitología
Patología
Química Médica
Reproducción
Virología

Física
Astronomía y Astrofísica
Cristalografía
Espectroscopia
Física Aplicada
Física Atómica, Molecular y Química
Física Matemática
Física Nuclear
Física, Estado Sólido
Física, Fluidos y Plasma
Física, Multidisciplinar
Física, Partículas y Campos
Geociencias, Multidisciplinar
Geología
Geoquímica y Geofísica
Meteorología y Ciencias Atmosféricas
Mineralogía
Oceanografía
Paleontología
Termodinámica

Ergonomía
Estudios de la Familia
Estudios de la Mujer
Estudios Étnicos
Estudios Medioambientales
Estudios por Áreas Geográficas
Ética
Ética Médica
Geografía
Geografía, Física
Historia de Ciencias Sociales
Medicina Alternativa
Ocio, Deporte y Turismo
Planificación y Desarrollo
Política Social y Servicios Sociales
Psicología
Psicología Aplicada
Psicología Biológica
Psicología Clínica
Psicología del Desarrollo
Psicología Educativa
Psicología Experimental
Psicología Matemática
Psicología Multidisciplinar
Psicología Social
Psicología, Psicoanálisis
Rehabilitación
Relaciones Empresariales y de Trabajo
Relaciones Internacionales
Servicios Médicos
Servicios y Política Sanitarios
Sociología
Temas Sociales
Urbanística

Humanidades

Arqueología
Arquitectura
Arte
Cine, Radio, Televisión
Danza
Estudios Asiáticos
Estudios culturales
Estudios Medievales y del Renacimiento
Filosofía
Folclore
Historia
Historia y Filosofía de la Ciencia

Ingeniería, Tecnología

Acústica
Ciencia de la Imagen y Tecnología Fotográfica
Ciencia de Materiales, Caracterización y Ensayos
Ciencia de Materiales, Cerámica
Ciencia de Materiales, Materiales Biológicos
Ciencia de Materiales, Materiales Compuestos
Ciencia de Materiales, Multidisciplinar
Ciencia de Materiales, Papel y Madera
Ciencia de Materiales, Revestimientos y Películas
Ciencia de Materiales, Textiles
Ciencia y Tecnología del Transporte
Control Remoto
Energía Nuclear
Energía y Combustibles
Informática, Aplicaciones Interdisciplinares
Informática, Cibernética
Informática, Hardware
Informática, Ingeniería del Software
Informática, Inteligencia Artificial
Informática, Sistemas de Información
Informática, Teoría y Métodos
Ingeniería Aeroespacial
Ingeniería Civil
Ingeniería de Fabricación
Ingeniería del Petróleo
Ingeniería Eléctrica y Electrónica
Ingeniería Geológica
Ingeniería Industrial
Ingeniería Marina
Ingeniería Mecánica
Ingeniería Medioambiental
Ingeniería Oceánica
Ingeniería Química
Ingeniería, Multidisciplinar
Instrumentación
Matemática e Informática Biológica
Mecánica
Metalurgia e Ingeniería Metalúrgica
Minería
Nanociencia y Nanotecnología
Óptica
Robótica
Sistemas de Automatización y Control
Tecnología de la Construcción
Telecomunicaciones
Transportes

Medicina Clínica

Humanidades, Multidisciplinar
Lenguaje y Lingüística
Lingüística
Literatura
Literatura Africana, Australiana, Canadiense
Literatura Alemana, Holandesa, Escandinava
Literatura Americana
Literatura Clásica
Literatura de las Islas Británicas
Literatura Eslava
Literatura Romance
Música
Poesía
Religión
Revisiones Literarias
Teatro
Teoría y Crítica Literarias

Matemáticas
Estadística y Probabilidad
Investigación Operativa y Ciencias de la Administración
Lógica
Matemáticas
Matemáticas Aplicadas
Matemáticas, Aplicaciones Interdisciplinares

Multidisciplinar
Ciencias Multidisciplinares
Educación, Disciplinas Científicas

Química
Electroquímica
Polímeros
Química Analítica
Química Aplicada
Química Física
Química Inorgánica y Nuclear
Química Orgánica
Química, Multidisciplinar

Alergia
Andrología
Anestesiología
Atención primaria
Audiología y Patología del habla y el lenguaje
Cirugía
Corazón y Sistema Cardiovascular
Dermatología
Drogodependencias
Enfermedades Infecciosas
Enfermedades Vasculares Periféricas
Gastroenterología y Hepatología
Geriatría
Gerontología
Hematología
Informática Médica
Ingeniería Biomédica
Medicina de Urgencia
Medicina Deportiva
Medicina Forense
Medicina Intensiva
Medicina Interna y General
Medicina Tropical
Medicina, Técnicas de Laboratorio
Neumología
Neuroimagen
Neurología Clínica
Nutrición y Dietética
Obstetricia y Ginecología
Odontología y Estomatología
Oftalmología
Oncología
Otorrinolaringología
Pediatría
Psiquiatría
Radiología y Medicina Nuclear
Reumatología
Salud Pública, Medioambiental y Laboral
Toxicología
Trasplantes
Traumatología y Ortopedia
Urología y Nefrología

Anexo II. Fundamentos teóricos del HJ-Biplot.

Los Biplot propuestos inicialmente por Gabriel (1971), JK-Biplot y GH-Biplot, son capaces de reproducir los elementos originales de una matriz de datos a través del producto interno de los marcadores fila y columna, mientras que el HJ-Biplot propuesto por Galindo (1986) ofrece la ventaja de ser una representación simultánea en sentido estricto. El JK-Biplot presenta máxima calidad de representación para los marcadores fila, mientras que en el GH-Biplot los marcadores columna son representados con máxima calidad, no así las filas. En base a estas características, Greenacre (1984) designó al JK-Biplot como RMP (*row-metric preserving*) y al GH-Biplot como CMP (*column-metric preserving*). En cambio, el HJ-Biplot propuesto por Galindo (1986) proporciona una representación simultánea donde marcadores fila y columna presentan idéntica bondad de ajuste por lo es que posible interpretar no sólo la posición de las filas y de las columnas, sino también las relaciones fila-columna.

Para lograr la transformación necesaria que permita establecer estos marcadores, los Biplot clásicos se basan en la aproximación de una matriz de datos X_{nxp} (con n individuos y p variables) por una de menor rango q, siendo $q < r$ mediante la descomposición en valores singulares (DVS) de X. A continuación, se lleva a cabo una factorización en matrices de marcadores fila y columna de forma que el producto escalar entre los marcadores aproxime de la mejor manera posible los valores de partida de X. Debido a que hay múltiples formas de factorización de matrices, las representaciones Biplot tendrán distintas propiedades en función de la métrica seleccionada. A continuación se exponen los fundamentos para la construcción de un HJ-Biplot. Se ha tomado como modelo la demostración presentada en Galindo, Barrera-Mellado, Fernández-Gómez y Martin (1996).

Galindo (1986) propone el HJ-Biplot como una representación gráfica multivariante de los datos de una matriz X_{nxp}, mediante marcadores $j_1, \dots, j_n$ para las filas y $h_1, \dots, h_p$ para las columnas, elegidos de forma que ambos marcadores puedan ser superpuestos en un mismo sistema de referencia con máxima calidad de representación.

Dada la DVS de la matriz X

$$X = U\ D\ V^T$$
$$\text{Con } U^T\ U = I \text{ y } V'V = I$$

es posible obtener una representación HJ-Biplot si se toma $J = U\ D$ como marcadores para las filas de X, y $H = V\ D$ como marcadores para las columnas. Esta elección de marcadores va a permitir, por un lado, que ambos marcadores sean representados en un mismo sistema de referencia y, por otro lado, que tanto filas como columnas presenten idéntica bondad de ajuste.

Si se tiene en cuenta que V son los vectores propios de $X^T X$ y que U está compuesta por los vectores propios de XX' y las relaciones que unen a U y V es posible escribir:

$$U = X\,V\,D^{-1}$$
$$V = X^T U\,D^{-1}$$

De forma que:

$$U\,D = X\,V\,D^{-1}\,D = X\,V$$
$$D\,V = D\,D^{-1}\,X^T\,U = X^T\,U$$

Ahora si a $X^T\,U$ se le denomina como B y a $X\,V$ como A se verifica que

$$X^T\,U = X^T X\,V\,D^{-1}$$
$$X\,V = X\,X^T\,U\,D^{-1}$$

Siguiendo la denominación empleada

$$B = X^T\,A\,D^{-1}$$
$$A = X\,B\,D^{-1}$$

Esto resulta en que la *h*-ésima coordenada de la columna *j*-ésima puede ser expresada según las coordenadas de las *n* filas mediante la siguiente expresión:

$$b_{jh} = (1/\alpha_h)\,\{x_{1j}\,a_{1h} + \cdots + x_{nj}\,a_{nh}\}$$

Si además la *h*-ésima coordenada de la fila *i*-ésima se expresa como una función de las *h*-ésimas coordenadas de las *p* variables:

$$a_{ih} = (1/\alpha_h) = \{x_{i1}\;b_{1h} + \cdots + x_{ip}\,b_{ph}\}$$

De esta forma, cada punto correspondiente a un individuo va estar situado en el baricentro de los puntos correspondientes a cada una de las variables, empleando como masa el valor que dicho individuo tome para cada una de esas variables. Por consiguiente, cuando un individuo tome un valor alto para una variable concreta, el punto que representa al individuo va a estar más cercano al punto que represente a esa variable, que al resto de variables.

Además, las dispersiones de ambas nubes de puntos se pueden aproximar mediante[33]

[33] Ambas aproximaciones están referidas a los valores propios contenidos en D^2.

$$X\,X^T = U\,D^2\,U^T \text{ para las columnas}$$
$$X'\,X = V\,D^2\,V^T \text{ para las filas}$$

Tanto la nube de puntos de las filas como la nube de puntos de las columnas presentan la misma dispersión y pueden relacionarse mediante combinaciones lineales simétricas.

Por otro lado, teniendo en cuenta las descomposiciones

$$X\,X^T = U\,D^2\,V^T$$
$$X'X = V\,D^2V^T$$

y eligiendo como marcadores

$$X\,X' = (U\,D)\,(D\,U^T)$$
$$X'X = (V\,D)\,(D\,V^T)$$

es posible representar las filas y columnas de las matrices $X\,X'$ y X^TX de igual forma que para la matriz X. En ambos casos, los marcadores son proyectados con la misma varianza (λ) y coinciden con los marcadores de las filas y las columnas de la representación HJ-Biplot de X. Por tanto, se garantiza que tanto individuos como variables están representados en un plano Biplot con la misma, y máxima, calidad de representación:

$$\frac{\lambda_1\lambda_2}{\sum_{i=1}^{r}\lambda_i}$$

Referencias

Gabriel, K.R. (1971). The biplot graphic display of matrices with application to principal component analysis. *Biometrika*, *58*(3), 453-467.

Galindo, M.P. (1986). Una alternativa de representación simultánea: HJ-Biplot. *Qüestioó*, *10*(1), 13-23.

Galindo, M.P., & Cuadras, C. (1986). *Una extensión del método Biplot y su relación con otras técnicas*. Publicaciones de Bioestadística y Biomatemática, 17. Barcelona: Universidad de Barcelona.

Galindo, M.P., Barrera-Mellado, I., Fernández-Gómez, M.J., & Martín-Casado, A.M. (1996). Estudio comparativo de ordenación de comunidades ecológicas basado en técnicas factoriales. *Mediterránea. Serie de Estudios Biológicos, Época* II, *15, 55-61*.

Greenacre, M.J. (1984). *Theory and applications of correspondence analysis*. London: Academic Press.

Anexo III. Fundamentos teóricos del Biplot Canónico.

El propósito del Biplot Canónico (Vicente-Villardón, 1992) o Manova-Biplot (Gabriel, 1972) es construir una representación simultánea de filas (grupos) y columnas (variables) de la matriz X donde los grupos estén separados por el máximo poder discriminante entre ellos. Siguiendo los trabajos de Vicente-Villardón (1992) y Varas, Vicente-Tavera, Molina y Vicente-Villardón (2005) se describen a continuación los fundamentos básicos para la obtención del Biplot Canónico empleado en esta tesis doctoral.

Supongamos que una matriz X de dimensión (nxp) (centrada), los n individuos pueden ser agrupados en K grupos claramente diferenciados con n_k individuos en cada uno de ellos $(k = 1,2, \dots, K; n = n_1 + n_2 + \cdots + n_k)$.

Sea Z_{nxk} una matriz que contiene variables indicadoras de cada uno de los grupos.

$$Z = \begin{pmatrix} 1 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & \dots & 0 & 1 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & \dots & 0 & 1 \end{pmatrix}$$

Como en este Biplot se trata de obtener una representación simultánea de las filas (centroides) y de las columnas de la matriz:

Se define

$$\overline{X} = S_{11}^{-1} Z'X,$$

la matriz que contiene las medias de los grupos para cada una de las variables consideradas, centradas respecto a la media global.

Con $S_{11} = Z' Z = diag(n_1 n_2, \dots, n_k)$,

la matriz de covarianzas dentro de los grupos $S = \frac{1}{n-k} (X'X - \overline{X'} S_{11} \overline{X})$

y la matriz de covarianzas entre los grupos $B = \frac{1}{\mathrm{k}-1} \overline{X'} S_{11} \overline{X}$ empleadas en el Análisis Canónico de Poblaciones.

Debido a las distintas unidades de medida de las variables y a la dispersión de los individuos que integran cada grupo, es necesario introducir una ponderación en relación a la matriz de covarianzas dentro de los grupos, y otra con respecto a los tamaños muestrales para tener así tener en cuenta que la precisión de las medias depende del tamaño muestral.

A partir de la descomposición en valores singulares (DVS) de la matriz

$$\overline{Y} = S_{11}^{1/2}\,\overline{X}\,S^{-1/2}$$

en la forma $Y = P\,D\,Q'$, donde las columnas de P están formadas por los vectores propios de $\overline{Y}\,\overline{Y}'$ y las columnas de Q por los vectores propios de $\overline{Y}'\overline{Y}$. Si se despeja $\overline{X}$ y se sustituye $\overline{Y}$ en la ecuación se obtiene que la matriz de medias $\overline{X}$ es de la forma:

$$\overline{X} = S_{11}^{-1/2}\,Y\,S^{1/2} = S_{11}^{-1/2}\,P\,D\,Q'S^{1/2}$$

A continuación, se pueden obtener representaciones canónicas Biplot que tendrán distintas propiedades en función de la DVS de la matriz $\overline{X}$. El GH-Biplot o CMP-Biplot no interesa en este caso debido a que los grupos están mal representados, mientras que el HJ-Biplot aunque presenta máxima calidad de representación para filas y columnas y es capaz de discriminar entre grupos a partir de los ejes factoriales, no cumple la propiedad de reproducir los elementos de la matriz original. En cambio, el JK-Biplot o RMP-Biplot ofrece las propiedades necesarias para la construcción de un Biplot Canónico (buena calidad de representación para los grupos y la capacidad de reproducir los elementos de la matriz de partida).

Se puede obtener un JK-Biplot a partir de la DVS generalizada de la matriz de medias mediante la siguiente selección de marcadores:

$J = S_{11}^{-1/2}\,P\,D_{(s)}$, que son los marcadores para las medidas de los g grupos (filas) y $K = S^{1/2}\,Q_{(s)}$, los marcadores de las p variables (columnas)

Esto proporciona una representación Biplot con las siguientes propiedades: (1) si los grupos son proyectados en el gráfico factorial, las coordenadas del primer eje representan la combinación lineal de variables que produce la mayor F de Snedecor univariante en el ANOVA (coordenadas discriminantes), (2) la distancia Euclídea entre dos marcadores de medias aproxima la distancia de Mahalanobis entre los grupos (magnitud del efecto) y (3) permite situar regiones de predicción en el gráfico factorial.

La bondad ajuste de las predicciones se mide a través de la varianza de cada una de las variables explicada por los factores (calidad de representación de las variables). La calidad de representación (bondad de ajuste de las filas), se puede estimar como:

$$\frac{\sum_{i=1}^{s} \lambda_i^2}{\sum_{i=1}^{r} \lambda_i^2}$$

Referencias

Gabriel, K.R. (1972). Analysis of meteorological data by means of canonical decomposition and biplots. *Journal of Applied Meteorology*, *11*, 1071–1077.

Varas, M.L., Vicente-Tavera, S., Molina, E., & Vicente-Villardón (2005). Role of Canonical Biplot method in the study of building stones: an example from Spanish monumental heritage. *Environmetrics*, *16*, 1-15.

Vicente-Villardón J.L. (1992). *Una alternativa a los métodos factoriales clásicos basada en una generalización de los métodos biplot*. PhD thesis. Salamanca, Spain: University of Salamanca.

Anexo IV. Índices de calidad de las publicaciones aportadas.

En este anexo se muestran los principales indicios de calidad de los artículos que conforman esta tesis doctoral realizada mediante compendio de publicaciones. Los cuatro artículos han sido publicados en revistas científicas internacionales incluidas en la base de datos *Web of Science*.

En la tabla I se presenta un resumen de la visibilidad y posición que ocupa cada revista en el *Journal Citation Reports*. Los cuatro artículos han sido publicados en revistas indexadas en la categoría *Information Science & Library Science*. El Capítulo 6 fue publicado en la *Revista Española de Documentación Científica (REDC)* que edita el CSIC y difunde, principalmente, trabajos sobre medición de la actividad científica y es la revista española con mayor visibilidad a nivel internacional en este ámbito. El Capítulo 7, que se encuentra en proceso de publicación, y el Capítulo 8 pueden hallarse en *Journal of Informetrics (JoI),* revista publicada por *Elsevier* dedicada a los aspectos matemáticos y estadísticos de los datos bibliométricos y a la aplicación de metodologías de análisis novedosas. La última contribución (Capítulo 9) ha sido publicada en *Journal of the Association for Information Science & Technology (JASIST)* que edita *Wiley*, revista que cubre una amplia gama de temas relacionados con la información y la tecnología y cuenta con una gran tradición a nivel internacional.

Tabla I. Indicios de calidad de las publicaciones.

	Revista	**Año**	**Vol./Nº**	**FI 2013**	**Ránking**	**Cuartil**
Capítulo 6	Revista Española de Documentación Científica	2013	36 (1)	0,717	40/84	Q2
Capítulo 7	Journal of Informetrics	En prensa	-	3,580	4/84	Q1
Capítulo 8	Journal of Informetrics	2015	9	3,580	4/84	Q1
Capítulo 9	Journal of the Association for Information Science & Technology	2014	65 (9)	2,230	9/84	Q1

Además, durante este período se ha publicado otro artículo en la revista *Scientometrics*, con la aplicación del HJ-Biplot para explorar el desempeño científico de los centros de investigación biomédica en red, y se ha participado en las principales conferencias internacionales del campo bibliométrico: *International Society of Scientometrics and Informetrics Conference (ISSI)* y *Science & Technology Indicators (STI)*. A nivel nacional, la metodología introducida en el Capítulo 9 para el análisis de los agradecimientos fue premiada como la mejor comunicación derivada de una tesis doctoral en el congreso *La colaboración científica: una aproximación multidisciplinar* (Valencia, 2014).

Diaz-Faes, A.A., & Bordons, M. (2013). A text mining approach exploring acknowledgements of papers. En: J. Gorraiz, E. Schielbel, C. Gumpenberger, M. Hörlesberger, H. Moed (Eds.), *Proceedings of the 14th International Society of Scientometrics and Informetrics Conference* (ISSI 2013) (pp. 2162-2164). Viena: ISSI. http://www.issi2013.org/Images/ISSI_Proceedings_Volume_II.pdf

Díaz-Faes, A.A., Galindo, M.P., & Bordons, M. (2014). Exploring internationality and collaborative behaviour of scientists in Social Sciences and Humanities. *Proceedings of the Science and Technology Indicators Conference 2014* (pp 164-168). Leiden (Holanda). Universidad de Leiden: CWTS.

Díaz-Faes, A.A., Galindo, M.P., & Bordons, M. (2013). Nuevas aproximaciones metodológicas a la colaboración científica: el análisis de los Agradecimientos. En: G. González-Alcaide, J. Gómez-Ferri, V. Agulló-Calatayud (Eds.) *La colaboración científica: una aproximación multidisciplinar* (pp. 269-280). Valencia (España): Nau Libres. ISBN-13: 9788476429303. *Premio a la Comunicación de Excelencia derivada de la realización de una Tesis Doctoral.*

Bordons, M., González-Albo, B. & Díaz-Faes, A.A. (2013). Colaboración científica e impacto de la investigación. En: G. González-Alcaide, J. Gómez-Ferri, V. Agulló-Calatayud (Eds.) *La colaboración científica: una aproximación multidisciplinar* (pp. 169-181). Valencia (España): Nau Libres. ISBN-13: 9788476429303.

Morillo, F., Díaz-Faes, A.A., González-Albo, B., & Moreno, L. (2014). Do networking centres perform better? An exploratory analysis in Psychiatry and Gastroenterology/Hepatology in Spain. *Scientometrics, 98* (2), 1401–1416. doi:10.1007/s11192-013-1183-5.